JMIR Publications
The Leading eHealth Publisher

The Use of Large Language Models Tuned with Socratic Methods on the Impact of Medical Students' Learning: A Randomised Controlled Trial

Cai Ling Yong, Mohammad Shaheryar Furqan, James Wai Kit Lee, Andrew Makmur, Ragunathan Mariappan, Clara Lee Ying Ngoh, Kee Yuan Ngiam

Submitted to: Journal of Medical Internet Research
on: March 03, 2024

Table of Contents

The Use of Large Language Models Tuned with Socratic Methods on the Impact of Medical Students' Learning: A Randomised Controlled Trial

Cai Ling Yong[1*]; Mohammad Shaheryar Furqan[2*] PhD; James Wai Kit Lee[3*] MBBS; Andrew Makmur[4*] MBBS; Ragunathan Mariappan[2*] MSc; Clara Lee Ying Ngoh[5*] MBChB; Kee Yuan Ngiam[2] MBBS

[1]Yong Loo Lin School of Medicine National University of Singapore Singapore SG
[2]Division of Biomedical Informatics Department of Surgery National University of Singapore Singapore SG
[3]Department of Surgery National University Health System Singapore SG
[4]Department of Diagnostic Imaging National University Health System Singapore SG
[5]Department of Medicine National University Health System Singapore SG
[*]these authors contributed equally

Corresponding Author:
Cai Ling Yong
Yong Loo Lin School of Medicine
National University of Singapore
10 Medical Dr, Singapore 117597
Singapore
SG

Abstract

Background: Large Language Models (LLM) are AI models that can generate conversational content based on a trained specified source of information (corpus).

Objective: The aim is to use these corpus-trained LLMs to limit the content offered by LLM, then using prompt engineering to teach using Socratic methods.

Methods: Two chatbots were created and deployed, powered by OpenAI’s GPT-3.5 model, with a medical-school textbook corpus. The first chatbot generates a brief summary and open-ended question. The second chatbot generates a case vignette from its pre-trained clinical cases, prompting users for a diagnosis. Both chatbots reply to the user’s response, commenting on the accuracy and asks further questions to encourage critical thinking. A randomised controlled trial was conducted on two groups comprising third year medical students. One group used both chatbots for 10 minutes while the other read the medical textbook. A 15-question test was administered to both groups before and after the intervention.

Results: Forty students participated in the study. The average of the group before and after reading the textbook (n=20) are 3.9 +/- 1.0 and 7.6 +/- 1.5 respectively (p<0.001). The average of the group before and after using the bot (n=20) are 3.9 +/- 0.9 and 12.8 +/- 1.6 respectively (p<0.001). The respective increase in results was 3.7 and 8.9.

Conclusions: Medical students’ learning showed a better performance using a LLM based chatbot compared to self-reading of medical information assessed using a standardised test. More studies are required to determine if LLM-based pedagogical methods are superior to standard education.

(JMIR Preprints 03/03/2024:57995)
DOI: https://doi.org/10.2196/preprints.57995

✓ **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

ORIGINAL ARTICLE

A RANDOMISED CONTROLLED TRIAL TO STUDY THE USE OF LARGE LANGUAGE MODELS TUNED WITH SOCRATIC METHODS ON THE IMPACT OF MEDICAL STUDENTS LEARNING

(Short title: RCT of LLM in Medical Education)

Cai Ling **Yong**[1], Mohammad Shaheryar **Furqan**, PhD[2], James Wai Kit **Lee**, MBBS[3], Andrew **Makmur**, MBBS[4], Ragunathan **Mariappan**, MSc[2], Clara Lee Ying **Ngoh**, MB ChB[5], Kee Yuan **Ngiam**, MBBS[2,3]

[1] Yong Loo Lin School of Medicine, National University of Singapore, Singapore

[2] Division of Biomedical Informatics, Department of Surgery, Yong Loo Lin School of Medicine, National University of Singapore, Singapore

[3] Department of Surgery, National University Health System, Singapore

[4] Department of Diagnostic Imaging, National University Health System, Singapore

[5] Department of Medicine, National University Health System, Singapore

Address for correspondence:

Miss Yong Cai Ling *(ORCID ID: 0009-0009-8651-868X)*
National University of Singapore, 10 Medical Dr, Singapore 117597
Tel.: +65 88096809 Email: yongcailing@u.nus.edu

All authors have seen and approved the manuscript
Declarations for each author: none for all the authors
Off-label or investigational use: none
Conflict of interest: none for all the authors

Word count of main text: 1992
Word count of abstract: 250

Number of references: 18
Number of tables: 2
Number of figures: 0

Keywords: Artificial Intelligence, Medical Education, Large Language Models, Medical Student, Chatbot

ABSTRACT

Introduction: Large Language Models (LLM) are AI models that can generate conversational content based on a trained specified source of information (corpus). The aim is to use these corpus-trained LLMs to limit the content offered by LLM, then using prompt engineering to teach using Socratic methods.

Methods: Two chatbots were created and deployed, powered by OpenAI's GPT-3.5 model, with a medical-school textbook corpus. The first chatbot generates a brief summary and open-ended question. The second chatbot generates a case vignette from its pre-trained clinical cases, prompting users for a diagnosis. Both chatbots reply to the user's response, commenting on the accuracy and asks further questions to encourage critical thinking. A randomised controlled trial was conducted on two groups comprising third year medical students. One group used both chatbots for 10 minutes while the other read the medical textbook. A 15-question test was administered to both groups before and after the intervention.

Results: Forty students participated in the study. The average of the group before and after reading the textbook (n=20) are 3.9 +/- 1.0 and 7.6 +/- 1.5 respectively (*P*<.001). The average of the group before and after using the bot (n=20) are 3.9 +/- 0.9 and 12.8 +/- 1.6 respectively (*P*<.001). The respective increase in results was 3.7 and 8.9. **Conclusion:** Medical students' learning showed a better performance using a LLM based chatbot compared to self-reading of medical information assessed using a standardised test. More studies are required to determine if LLM-based pedagogical methods are superior to standard education.

BACKGROUND

Artificial intelligence (AI) was created with the aim of emulating the problem-solving and decision-making capabilities of human intelligence through the use of technology. The birth of the AI conversation in 1950 by pioneer Alan Turing sparked much debate, yet the release of OpenAI’s ChatGPT in 2022 made the applications of Generative AI undeniably ubiquitous. Across the world, consumers and companies alike have begun using AI to automate menial tasks, as well as innovate creative content. Similarly, AI has been explored in almost every field of medicine.[1] For instance, AI is aiding ophthalmologists in the early diagnosis and treatment of conditions such as diabetic retinopathy with separate innovations of the Topcon NW400 fundus camera and trained Convoluted Neural Networks from databases of retinal images.[2, 3] Among these notable advancements in AI, Large Language Models (LLMs) stand out for its ability to recognise and generate text.

Large Language Models (LLMs) are the algorithmic basis for chatbots such as OpenAI’s ChatGPT and its more recent model, GPT-4. LLMs are large pre-trained AI models that can be easily repurposed across a wide range of domains. Fundamentally, LLMs leverage deep neural networks, which are complex structures with layers of statistical correlation, that enable complex information abstraction.[4] The capacity of these interactive models to extract knowledge embedded in medical corpora at scale holds considerable promise.[5]

LLMs can play a pivotal role in medical education, particularly in the generation of virtual simulated patients or quizzes for medical students.[6] During the curriculum, these use cases can be a useful way for students to learn the fundamentals on medicine while enhancing their learning of basic Entrustable Professional Activities.[7]

Critical thinking is an essential competency in the medical curriculum, required for managing medical ambiguity to treat illnesses. The concept of critical thinking is thought to have originated from the teachings of classic Athenian philosopher Socrates. Socratic questioning refers to the process of asking a series of specific, targeted inquiries of increasing depth and difficulty to allow students to understand the limits of their knowledge. Socratic questioning has been shown to be effective in developing critical thinking in healthcare students.[8]

We conducted a randomised controlled trial (RCT) on third year medical students at the start of their General Surgery rotation in the Yong Loo Lin School of Medicine. The primary objective of the study is to investigate how combining LLMs with Socratic methods can be an innovative avenue to enhance medical education. We hypothesize that the use of LLMs guided by Socratic questioning techniques can produce better learning outcomes compared to conventional self-study approaches.

METHODS

Creation and Deployment

Two Telegram chatbots were created for the purpose of this study. Telegram is a free, cross-platform, cloud-based messaging app.[9] It is used as the unofficial channel for dissemination of information and communication for the batch of third year medical students. Both chatbots were created and deployed on Amazon Web Services (AWS) Lambda. The LLMs were powered by OpenAI's GPT-3.5 model. The reference corpus was limited to a school-approved textbook. Few-shot learning was performed with reference clinical cases. Few-shot learning refers to a set of machine learning methods that learn how to complete tasks with a small number of labelled training examples.[10] The function of the chatbots are to generate questions to test the users knowledge, which was achieved through prompt engineering.

The first chatbot was programmed using prompt engineering to respond to a topic given by the user, however generic or specific, through generating a brief summary of the topic and ending with an open-ended question. The user will then respond to the question. The chatbot replies to the user's response, commenting on the accuracy, providing explanations, and asks further questions to encourage critical thinking. Sample interactions of the user and the first chatbot are shown in Table 1.

The second chatbot was programmed using prompt engineering to generate a case vignette from its pre-trained clinical cases in response to a topic given by the user. In the same message, it will also ask the user for a diagnosis. Following the user's response, the bot will comment on the accuracy and provide explanations. It will then continue to ask further open-ended questions similar to the first chatbot to encourage critical thinking. Sample interactions of the user and the second chatbot are shown in Table 2.

Sampling and Recruitment

The randomised controlled trial (RCT) was conducted over a 2-week period at the Yong Loo Lin School of Medicine, National University of Singapore. The inclusion criteria was actively enrolled third year medical students during the first 2 weeks of their General Surgery rotation. Participants were excluded if did not meet the above requirements, or had prior exposure to General Surgery content.

The medical students were invited to participate in the study via a telegram message sent to the class chat group. Each participant was then randomly allocated into the control or experimental group to ensure equitable distribution. Prior to beginning the trial, informed consent was obtained from every participant.

Assessment

Before each intervention, a 15-question test conprising 14 Multiple Choice Questions (MCQs) and one case vignette diagnosis question was administered to both groups, and scored according to their responses. Each question carries a weight of 1. The test was designed to establish a baseline knowledge level of all participants. The scores of participants were recorded. The correct answers were not divulged to the students.

Subsequently, Group 1 (Control) was asked to read the softcopy of the medical textbook that is similar to that used to train the chatbot for 10 minutes. Group 2 (Experimental) was asked to use both chatbots for 10 minutes, actively responding to the prompts.

Following the interventions, both groups were administered the same 15 question test. Scores were recorded without divulging the answers, allowing for a comparison on knowledge acquisition over the 10 minutes.

Analysis

A paired t-test was performed to check for statistical significance of the changes in test scores after each intervention. A threshold of *P*<.05 was established to consider the change statistically significant.

The study's statistical power and requisite sample size was calculated using a study power calculator. Parameters were as follows: Power = 80%, alpha = 0.05, anticipated means for Groups 1 and 2 respectively = 7 +/- 2 and 10, enrollment ratio = 1. The result was a requisite enrollment of 7 for each group.

RESULTS

A total of forty students participated in the study, with twenty students each in the control and experimental groups.

The average scores for the control group (n=20) before reading the textbook was 3.9 +/- 1.0 with range of 2 to 6. Their average scores after reading the textbook for 10 minutes was 7.6 +/- 1.5 with range of 5 to 11.

The average scores for the experimental group (n=20) before using the bot was 3.9 +/- 0.9 with range of 2 to 5. Their average scores after reading the textbook for 10 minutes was 12.8 +/- 1.6 with range of 9 to 15.

The respective increase in results for the control and experimental group respectively was 3.7 and 8.9.

Statistical Analysis

A paired t-test was used to compare statistical significance. Comparing the increase in scores for both groups before and after each intervention, the p scores were <0.001 for both.

Feedback

Feedback on the chatbots was gathered from some participants in the experimental group after using the chatbots. The responses can be categorised into the following themes:

Potential for Use

Participants expressed that the chatbot was an innovative solution that encourages critical thinking.

“I think it is very good at prompting you to think by asking you questions. It also provides good answers to my questions.”

Some expressed that the chatbots held great promise.

“This has a lot of potential. Very exciting.”

Areas for Improvement

Some participants expressed that after rigorous use, the questions became repetitive.

“Sometimes the question it asked is a bit repetitive. For example, it asked me the pros and cons of laparoscopic appendectomy 3 times.”

Occasionally, the bot will not ask a question following the explanation. Further prompting by the user by asking “Ask me another question” often resolves this issue.

DISCUSSION

Principal Findings

AI chatbots are widely used in a variety of industries, with increasing prominence in the healthcare industry.[11] However their use in medical education has been largely unexplored. This quantitative study explored the utility of socratic questioning LLM chatbots in medical education. Compared to the group using just textbooks, our study demonstrated improved test results in the group that used the LLM chatbot to learn new content.

The findings of this study have important implications. First and foremost, they underscore the advantages of LLM chatbots in medical education. The rapid advancements in the health-tech industry necessitates the incorporation of technology into medical education. Education technology should be embraced for the ways it can contribute to and support learning, as well as the acquisition and maintenance of clinical expertise.[12] Our findings support this view and suggests that the novel use of LLM chatbots in medical education are beneficial to learning. It remains to be seen if LLM will be able to effectively teach the higher aspirant rungs in Miller’s pyramid, such as clinical reasoning and diagnostic acumen.[13] Traditionally, medical schools have had to direct significant resource and healthcare utilisation, in order to recreate the settings for students in their clinical years to learn clinical reasoning. Another criticism of the traditional model is that it was challenging to recreate homogenous clinical experiences for all students. If successful, this will create a large value proposition for the incorporation of LLM into medical school curriculum.

Beyond just the academic supplementation that LLM chatbots can provide, the integration of LLMs in the curriculum will make undergraduate medical students more receptive to the use of technology in their future careers as medical practitioners. The current perceived awareness and engagement of AI among doctors is low across various healthcare systems such as the NHS in the United Kingdom, among Syrian doctors and medical students, and fellow ASEAN nation Vietnam.[14-16] By introducing LLM early in the undergraduate medical curriculum, this may result in a shift in increased perception of the utility of AI in healthcare by future healthcare practitioners. This is essential to ensure that the next generation of physicians is prepared to embrace technology to aid diagnostic and treatment decisions.[17]

Limitations

Several limitations should be considered when interpreting the results of this study. Firstly, the study was conducted only on Year 3 students undertaking their GS rotation at one medical school, which may limit the generalisability of the findings to other populations or topics in medical school curricula. Similarly, the study was conducted on a medical school in Singapore, a city-state with a high smartphone penetration rate. This may limit the generalisability of the findings to regions with lower smartphone penetration rates.

Secondly, the study was designed to test only knowledge acquisition, other aspects such as long term knowledge retention or clinical application were not assessed. Thirdly, the sampling framework used was a convenience sample, a non-probability sampling method, which meant that participants were selected because of their availability. This implies a risk that individuals selected were not equitably targeted.[18]

Implications

The conclusions which were drawn from this study highlighted areas which require further research to confirm its findings. Firstly, more randomised controlled trials are warranted, by expanding the scope of future questionnaires to include more topics, and obtaining a larger sample size. Secondly, conducting a Focus Group Discussion with students, as well as stakeholders of the medical education community and LLM experts, will provide valuable insight into refining the model.

CONCLUSION

Two LLM chatbots incorporated with socratic questioning techniques were developed and trained using prompt engineering. A standardised test was administered to the students using the chatbots, and to those who used a textbook instead. This showed a better performance among students who used the LLM chatbots compared to those who only read the textbook. Future research should continue to explore if LLM-based pedagogical methods are superior to standard education.

CONFLICT OF INTEREST: None declared.

FINANCIAL SUPPORT: None declared.

AUTHOR CONTRIBUTIONS: All authors made substantial contributions to the conception and design of the work. All authors made contributions to drafting and critically revising the manuscript. All authors have read and approved the submitted version.

DATA AVAILABILITY STATEMENT: The data underlying this article will be shared on reasonable request to the corresponding author.

REFERENCES

1. Ramesh AN, Kambhampati C, Monson JR, Drew PJ. Artificial intelligence in medicine. Ann R Coll Surg Engl. 2004 Sep;86(5):334-8. PMID: 15333167. doi: 10.1308/147870804290.

2. Keskinbora K, Güven F. Artificial Intelligence and Ophthalmology. Turk J Ophthalmol. 2020 Mar 5;50(1):37-43. PMID: 32167262. doi: 10.4274/tjo.galenos.2020.78989.

3. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA. 2016;316(22):2402-10. doi: 10.1001/jama.2016.17216.

4. Brants T, Popat A, Xu P, Och FJ, Dean J, editors. Large language models in machine translation. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL); 2007.

5. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023 Aug;620(7972):172-80. PMID: 37438534. doi: 10.1038/s41586-023-06291-2.

6. Eysenbach G. The Role of ChatGPT, Generative Language Models, and Artificial Intelligence in Medical Education: A Conversation With ChatGPT and a Call for Papers. JMIR Med Educ. 2023 Mar 6;9:e46885. PMID: 36863937. doi: 10.2196/46885.

7. Safranek CW, Sidamon-Eristoff AE, Gilson A, Chartash D. The Role of Large Language Models in Medical Education: Applications and Implications. JMIR Med Educ. 2023 Aug 14;9:e50945. PMID: 37578830. doi: 10.2196/50945.

8. Oyler DR, Romanelli F. The fact of ignorance: revisiting the Socratic method as a tool for teaching critical thinking. Am J Pharm Educ. 2014 Sep 15;78(7):144. PMID: 25258449. doi: 10.5688/ajpe787144.

9. Iqbal MZ, Alradhi HI, Alhumaidi AA, Alshaikh KH, AlObaid AM, Alhashim MT, et al. Telegram as a Tool to Supplement Online Medical Education During COVID-19 Crisis. Acta Inform Med. 2020 Jun;28(2):94-7. PMID: 32742059. doi: 10.5455/aim.2020.28.94-97.

10. Ge Y, Guo Y, Das S, Al-Garadi MA, Sarker A. Few-shot learning for medical text: A review of advances, trends, and opportunities. J Biomed Inform. 2023 Aug;144:104458. PMID: 37488023. doi: 10.1016/j.jbi.2023.104458.

11. Lee P, Bubeck S, Petro J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. N Engl J Med. 2023 Mar 30;388(13):1233-9. PMID: 36988602. doi: 10.1056/NEJMsr2214184.

12. Han H, Resch DS, Kovach RA. Educational technology in medical education. Teach Learn Med. 2013;25 Suppl 1:S39-43. PMID: 24246105. doi: 10.1080/10401334.2013.842914.

13. Thampy H, Willert E, Ramani S. Assessing Clinical Reasoning: Targeting the Higher Levels of the Pyramid. J Gen Intern Med. 2019 Aug;34(8):1631-6. PMID: 31025307. doi: 10.1007/s11606-019-04953-4.

14. Ganapathi S, Duggal S. Exploring the experiences and views of doctors working with Artificial Intelligence in English healthcare; a qualitative study. PLoS One. 2023;18(3):e0282415. PMID: 36862694. doi: 10.1371/journal.pone.0282415.

15. Vuong QH, Ho MT, Vuong TT, La VP, Ho MT, Nghiem KP, et al. Artificial Intelligence vs. Natural Stupidity: Evaluating AI readiness for the Vietnamese Medical Information System. J Clin Med. 2019 Feb 1;8(2). PMID: 30717268. doi: 10.3390/jcm8020168.

16. Swed S, Alibrahim H, Elkalagi NKH, Nasif MN, Rais MA, Nashwan AJ, et al. Knowledge, attitude, and practice of artificial intelligence among doctors and medical students in Syria: A cross-sectional online survey. Front Artif Intell. 2022;5:1011524. PMID: 36248622. doi:

10.3389/frai.2022.1011524.

17. Stoeklé HC, Charlier P, Hervé C, Deleuze JF, Vogt G. Artificial intelligence in internal medicine: Between science and pseudoscience. Eur J Intern Med. 2018 May;51:e33-e4. PMID: 29428496. doi: 10.1016/j.ejim.2018.01.027.

18. Suen LJ, Huang HM, Lee HH. [A comparison of convenience sampling and purposive sampling]. Hu Li Za Zhi. 2014 Jun;61(3):105-11. PMID: 24899564. doi: 10.6224/jn.61.3.105.

Supplementary Files