Web3 Journal: ML in Health Science

ISSN 2942-8726

2024. Vol. 1(1)

Original Article

TxnHash: 0x78d63d05e8ba3c02420c7b04c1b466c4d9894850f6ad3fb0bab45d2aa70b3f4a

DOI: 10.62487/8ny3zh09

Prevention of Artificial Intelligence (AI) Misuse in Online Medical Education

Yury Rusinovich[1,2], Volha Rusinovich[1,2]

Abstract.

Aim: This study aims to assess the capabilities of artificial intelligence (AI) in answering online Continuing Medical Education (CME) courses to find the resistant to AI-misuse strategies. **Materials and Methods:** The study evaluated 30 CME online courses from popular American (ACCME), European (EACCME), and German Medical Association accredited online platforms, including Medscape, eaccme.uems.eu, Springer Nature, der-niedergelassene-arzt, and Aerzteblatt. ChatGPT Version 4.0 with integrated plugins for interactive AI chats with documents, web access to scientific databases, and interactive AI chats with videos was used to answer the CME evaluation questions. A special scoring system, referred to as "complexity score," was introduced in the study. This system has two main objectives: first, to assess strategies that prevent the misuse of AI in medical online education; second, to measure the effort that physicians must invest to answer CME questions using AI. **Results:** AI was used to answer a total of 248 questions, divided into three categories: ACCME accredited courses: 7 credits; EACCME accredited courses: 9.5 credits; German CME courses: 28 credits. AI successfully completed the quiz in 90% of cases (27 courses) and showed an accuracy rate of 86%. 213 out of 248 questions were correctly answered: 38 out of 48 ACCME questions; 85 out of 100 EACCME questions; 90 out of 100 CME questions. The outcome "AI error" was significantly associated only with a higher number of questions in the quiz: p-value 0.01. However, this predictor had no impact on the AI's ability to successfully complete the entire quiz. The AI failure rate was significantly associated with learning materials based on new studies without open access: p-value 0.02 and the need to view all learning materials to gain access to the quiz: p-value 0.02. A higher complexity score of the course was significantly associated with AI failure rates: p-value 0.0034. **Conclusion:** This study has shown that AI can successfully answer medical quiz questions even without access to learning materials. Therefore, the best strategy to prevent the misuse of AI in CME online training is to align human learning with AI feeding. Access to the quiz should only be possible after a complete review of the learning materials. This could be achieved by setting a fixed time or through multiple slides with separate access to each slide and subsequent quiz access.

Keywords: ChatGPT, AI in medical education, CME, Medical quizzes

Background:

Introduction to generative artificial intelligence (AI): Generative AI is still in its infancy. If we consider the development of AI as an S-curve, we are just at the beginning of the steep ascent. Python, a key programming language for deep learning, has been around since 1991[1]. Tensor-Flow, a very well-known software for computationally intensive tasks and machine learning, was released by Google in 2015[2]. The first so-called "Transformer," a special technology for AI, was also introduced by Google in 2017[3].

[1]ML in Health Science, Leipzig, Germany

[2]University Hospital Leipzig, Germany

Corresponding author: Yury Rusinovich

Email: info@mlhs.ink

OpenAI developed the first Generative Pre-trained Transformer (GPT) in 2018, a type of AI that can generate texts[4]. The number of these GPT models is constantly growing. Today, there are many different and powerful AI models that can understand and generate texts, such as ChatGPT-4 (OpenAI, San Francisco, USA), Bard (Google AI, San Francisco, USA), Llama 2 (Meta AI, New York, USA), Claude 2 (Anthropic, San Francisco, USA), and Stable Diffusion XL (Stability AI Ltd).

Just two years ago, Sam Altman, the CEO of OpenAI, said that in the coming years we will have AI-driven medical advisors who can provide high-quality medical advice to everyone[5]. He also spoke of AI teachers who can teach mathematics or other subjects[5]. Today, we know that AI can even pass medical state examinations[6,7].

The role of AI in online medical education.

The role of AI in online monitoring of medical education. On one hand, machine learning (ML) can enhance the quality of health data science as it can analyze large amounts of data quickly, independently, and automatically, providing precise, evidence-based answers. On the other hand, there's a risk that AI could be overused or misused by healthcare professionals and those in medical education. This could lead to a devaluation of human roles in medicine and halt medical progress. In a very negative scenario, humans might even relinquish full control to AI, which, according to an experiment called "AI Box," isn't so far-fetched[8,9].

The next issue with AI misuse in medical education is that we might no longer distinguish whether a human or a machine is doing work that directly impacts human life, as the algorithms developed so far are not sufficient for this purpose[10].

Hypothesis:

If we can no longer distinguish whether we are dealing with AI or a human in one of the most critical areas of medicine, namely medical education, there arises a serious risk that natural, human-centered medicine will perish. The abstract, human way of thinking, which contributes to medical advancements and more humane patient treatment, could be replaced by the extremely rational logic of machine learning or "The Zen of Python"[11].

In this scenario, we should develop strategies that promote collaboration between humans and machines and prevent the misuse of AI.

Aim:

This study aims to examine the performance of AI in the context of "off the job" medical personnel development to identify strategies for the prevention of AI misuse.

Material and Methods:

In the study, 30 CME online courses from popular American (ACCME), European (EACCME), and German Medical Association accredited online plat-forms, including Medscape, eaccme.uems.eu (WebSurg, Vielgesundheit, and Diabetes Symposium), Springer Nature (Springer Medicine), der-niedergelassene-arzt, and Aerzteblatt, were evaluated. To answer the CME evaluation questions, ChatGPT Version 4.0 with integrated plugins for interactive AI chats with documents (AIPDF), web access to scientific databases (Metaphor and ScholarAI), content access via provided links (Access Link), and interactive AI chats with videos (MixerBox ChatVideo) was used.

Predictors:

In the study, a special scoring system was introduced, referred to as the "complexity score." This system has two main objectives: Firstly, to assess strategies that prevent the misuse of Artificial Intelligence (AI) in medical online education. Secondly, to measure the effort that trainee physicians must invest to answer CME questions using AI. The complexity score includes three different approaches that consider a total of nine different predictors for further statistical evaluation (**Table 1**):

1. Alignment of human learning and AI feeding:

- The learning materials consist of multiple slides with individual links for each slide. Merging, text recognition, or separate analysis of each file increases the time needed to feed the AI with learning materials. The use of AIPDF is either not possible or the effort is as great as with human learning. Categorical variable "Apply AIPDF": yes/no.

- Necessary to go through all the learning materials to gain access to the quiz. Categorical variable: yes/no.

2. Restriction of AI's access to learning materials

- The learning materials consist of long videos or audio files that are only available through secured access. As a result, access for AI is either prohibited or data analysis is not possible. Categorical variable: yes/no.
- The amount of non-text-based information (flowcharts, tables, images) in the learning materials. The amount of non-text-based materials was assessed in relation to the total learning information, based on the CME score of the course. 1 CME point was considered 100%, and a non-textual unit was considered 1% of that. Continuous variable: %.
- Case-specific questions in the quiz. Categorical variable: yes/no.
- Learning materials based on studies without open access. Categorical variable: yes/no.

3. Confusion techniques:

- Numerical approximation values in the quiz. Categorical variable: yes/no.
- Elimination questions. Categorical variable: yes/no.
- Abbreviations. Categorical variable: yes/no

Outcome:

The outcomes (dependent variables) in the analysis were:

1. AI Error: The machine made a mistake in the quiz. Categorical variable: yes/no.
2. AI Failure Rate: The machine did not pass the quiz. Categorical variable: yes/no.

Complexity Score:

Predictors that had a significant impact on the outcome "AI Failure Rate" were rated with 2 points in the complexity score, while predictors without significant influence received only 1 point. In analyzing the outcome "AI Error," the number of questions in the quiz was also considered. Quizzes with more than 10 questions received an additional point in the calculation of the complexity score. The total score of each course served as a numerical variable for statistical analysis.

Statistics:

The interactions between the outcomes and predictors were determined using logistic regression analysis of the dataset and the likelihood ratio test. To avoid overfitting the model, the influences of the predictors were assessed separately. A Cook's distance of more than 0.5 was considered influential. A p-value of $p<0.05$ (Pr(>Chisq=)) was considered statistically significant. P-values were not corrected for multiple comparisons. Data collection was carried out using an Excel spreadsheet. R was used for the statistical analysis.

Results:

The AI was used to answer a total of 248 questions, which were divided into three categories:

- 10 ACCME accredited courses: 7 credits.
- 10 EACCME accredited courses: 9.5 credits.
- 10 German CME courses: 28 credits.

The AI successfully completed the quiz for obtaining credits in 90% of cases (27 courses) and showed an accuracy rate of 86% (213 out of 248 questions correctly answered): 38 out of 48 ACCME questions; 85 out of 100 EACCME questions; 90 out of 100 CME questions. The outcome "AI Error" was significantly associated only with a higher number of questions in the quiz: p-value 0.01. However, this predictor had no impact on the AI's ability to successfully complete the entire quiz. The AI failure rate was significantly associated with learning materials based on new studies without open access: p-value 0.02. Another variable significantly associated with this outcome was the need to view all the learning materials to gain access to the quiz: p-value 0.02. The direct amount of non-text-based learning material had no significant impact on the outcome. However, courses with a proportion of non-text-based learning materials > 20% were associated with higher failure rates: p-value

Variable	% (n)	P-value
The use of AIPDF	66%(20)	0.21
Courses with access to the quiz only after reviewing all the learning materials	3%(1)	0.026**
Amount of non-text-based learning materials in the courses	24.3%(23.7; 30)*	0.12
Courses with a proportion of non-text-based learning materials > 20%	53%(16)	0.04**
Courses with learning material as long video or audio files with secured access.	20%(6)	0.059
Courses with learning materials based on studies without open access	3%(1)	0.026**
Courses with case-specific questions in the quiz	36%(11)	0.27
Courses with numerical approximation values in the quiz	20%(6)	0.56
Courses with elimination questions	36%(11)	0.089
Courses with abbreviations in the quiz	86%(26)	0.34
Complexity score***	3.5(1.9; 30)	0.0034**

Table 1: Complexity score of the courses. Outcome "AI Failure Rate"

n - Number of courses.
* mean (SD, n).
** p-value from regression analysis. <0.05 - by conventional criteria, this difference is statistically significant
*** Each predictor with a significant impact on the outcome received 2 points, while those without a significant impact received only 1 point.

0.04. A higher complexity score of the course was significantly associated with "AI Failure Rates" and "AI Error": p-value < 0.005. There were no influential observations in the dataset according to Cook's distance.

Table 1 provides an overview of the complexity of the courses in terms of strategies to prevent AI misuse.

Discussion:

Practical standpoint:

This study has shown that AI can successfully solve medical quiz questions and pass online CME courses. The large language models currently integrated into ChatGPT have up-to-date databases and constant access to scientific libraries. Therefore, AI does not need access to learning materials to correctly answer the questions. However, easy access to learning materials, such as a PDF file, can facilitate the misuse of AI. Only new studies without open access can lead to AI failing the quiz. Other factors individually had no significant impact on this outcome. However, the combination of various strategies, such as increasing the non-textual content in learning materials, questions based on case reports, and various confusion techniques, can cause AI to fail.

Based on the results of this study, the best strategy to prevent the misuse of AI is to align learning and AI feeding. For example, access to the quiz should only be possible after a complete review of the learning materials. This could be achieved by setting a fixed time or through multiple slides with separate access to each slide and quiz access only after going through all the slides.

New strategies for aligning human learning and AI feeding should be implemented in the medical education sector to ensure high-quality medical education and prevent the medical community's dependence on AI.

Limitations:

The study has the following limitations:

1. The main limitation of this study is that it used an empirical model to assess the performance of AI and the required human effort. However, to the best of the author's knowledge, there are no sufficient literature data on this topic. This is the first study to use an original scoring system to assess strategies for preventing the misuse of AI in medical education.
2. It was not possible to match questions with errors, as not all quizzes allow the review of your answers. Only the number of correct answers and the overall result could be included in the analysis.

3. The standards for the required learning time and the number of credits earned varied between medical associations.

Conclusion:

This study has shown that AI can successfully answer medical quiz questions even without access to learning materials. Therefore, the best strategy to prevent the misuse of AI in CME online training is to align human learning with AI feeding. Access to the quiz should only be possible after a complete review of the learning materials. This could be achieved by setting a fixed time or through multiple slides with separate access to each slide and subsequent quiz access.

Conflict of Interest: The authors state that no conflict of interest exists.

Authorship: YR: Concept, data analysis, original draft. YR, VR: Review and editing.

Ethical Statement: The author confirms that all learning materials were thoroughly studied and understood before the application of artificial intelligence (AI). AI was used as a tool for the research process.

Use of Material: The material used in this manuscript was originally developed and submitted by YR as part of the submission work for the Master's program in Health Business Administration (MHBA).

References

1 Python Institute. Python® – the language of today and tomorrow. https://pythoninstitute.org/about-python. Accessed October 4, 2023.

2 Alec Mccabe. History and Basics of Tensorflow. https://medium.com/@alec.mccabe93/history-and-basics-of-tensorflow-eaee87c6aef0. Accessed October 4, 2023.

3 Jakob Uszkoreit. Transformer: A Novel Neural Network Architecture for Language Understanding. https://blog.research.google/2017/08/transformer-novel-neural-network.html?m=1. Accessed October 4, 2023.

4 Radford A, Narasimhan K, Salimans T, Sutskever I. Improving Language Understanding by Generative Pre-Training. https://openai.com/research/language-unsupervised. Accessed October 4, 2023.

5 Klein E. Transcript: Ezra Klein Interviews Sam Altman. https://www.nytimes.com/2021/06/11/podcasts/transcript-ezra-klein-interviews-sam-altman.html. Accessed October 4, 2023.

6 Friederichs H, Friederichs WJ, März M. ChatGPT in medical school: how successful is AI in progress testing? Med Educ Online. 2023;28(1):2220920. doi:10.1080/10872981.2023.2220920.

7 Gilson A, Safranek CW, Huang T, 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. doi:10.2196/45312.

8 Yudkowsky ES. Singularity: The AI-Box Experiment. https://www.yudkowsky.net/singularity/aibox. Accessed October 18, 2023.

9 Altman S. Machine intelligence, part 2. https://blog.samaltman.com/machine-intelligence-part-2. Updated March 2, 2015. Accessed October 5, 2023.

10 Kirchner JH, Ahmad L, Aaronson S, Leike J. New AI classifier for indicating AI-written text. https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text. Updated January 31, 2023. Accessed October 5, 2023.

11 Peters T. PEP 20 – The Zen of Python. https://peps.python.org/pep-0020/. Updated September 9, 2023. Accessed October 25, 2023.