Educating the Next Generation: Making Large Language Models Mandatory in Medical Training

Plagwitz Lucas*
Institute of Medical Informatics
University of Muenster
Muenster, Germany
lucas.plagwitz@uni-muenster.de

Fujarski Michael*
Institute of Medical Informatics
University of Muenster
Muenster, Germany
michael.fujarski@uni-muenster.de

Sandmann Sarah
Institute of Medical Informatics
University of Muenster
Muenster, Germany
sarah.sandmann@uni-muenster.de

Brix Tobias
Institute of Medical Informatics
University of Muenster
Muenster, Germany
tobias.brix@uni-muenster.de

Varghese Julian
Institute of Medical Informatics
University of Muenster
Muenster, Germany
julian.varghese@uni-muenster.de

***Abstract*—The deployment of Large Language models (LLMs) introduces significant opportunities in the medical domain. Starting from supporting medical documentation to clinical decision making, the applications appear boundless. However, this brings challenges such dissemination of existing societal biases or the generation of misinformation. To address a positive future for LLMs in medicine, balancing both opportunities and challenges, we have designed and implemented a new mandatory course for medical students as part of their core curriculum. It includes a 90-minute session divided into three parts: a technical overview of LLMs, a live Turing Test to experience and evaluate human versus machine-generated responses, and a practical group exercise to assess the application of LLMs in three different medical tasks. This paper discusses the course structure, release the implementation of our Turing Test and LLM communication interface, and provides an initial evaluation, illustrating the critical role of education and fostering responsible use of emerging LLM technology in healthcare and research.**

Keywords—Large Language Models, Medicine, Education, Turing Test, Artificial Intelligence, Course Plan

I. INTRODUCTION

The emergence of Large Language Models (LLMs) has already transformed the way humans interact with machines across various domains, leading to a future of increasingly sophisticated technological interactions. As an example, OpenAI's GPT-4 has not only mastered human and programming languages with remarkable accuracy but is also beginning to significantly influence fields such as medicine [1, 2]. Across a wide range of application, LLMs have shown impressive results. For instance, in clinical documentation, they excel at extracting information from clinical notes and generating medical letters [3]. Furthermore, our previous work has demonstrated their advanced capabilities in complex clinical decision support situations [4].

While these models exhibit exceptional capabilities, their application necessitates a profound understanding of their limitations, biases, ethical implications, and optimized usage through prompt engineering [5]. Blind reliance on LLMs without such awareness can raise unintended consequences, ranging from perpetuating societal biases to spreading misinformation. This imperative of informed usage is particularly crucial in the medical domain, where decisions directly impact human lives.

The advancing medical capabilities of LLMs as well as the challenges regarding biases and ethical implications highlight a strong need for education. It is crucial to familiarize healthcare professionals with the responsible utilization of LLMs. This includes fostering a basic understanding on how these models operate, their potential biases, and the necessity of substantiating their outputs with clinical expertise. Moreover, it involves internalizing a habit of critical verification, where practitioners actively question the accuracy of recommendations generated by LLMs. A balanced approach is necessary to neither demonize nor glorify this new technology, addressing both its opportunities and challenges.

The balancing of strengths and weaknesses as part of the education process is not novel. Research impressively demonstrates how targeted teaching can optimize these aspects [6, 7]. However, these studies primarily focus on the LLM as a tool for optimized teaching and less on teaching and experiencing the LLM itself. Nevertheless, a specialized introduction to its use is equally necessary for students. In this way, fields like medicine can explore and cultivate their own benefits through new generations of medical professionals. However, there is a notable lack of specific guidance on how to familiarize medical students with LLMs.

Hence, we developed a course plan which addresses the aforementioned issues by teaching both, the fundamentals of

* These authors contribute equally to this work.

LLMs as well as the integration of such systems into clinical reasoning. We implemented a concise 90-minute course obligatory for every advanced medical student at our university in the ninth semester. The course is divided into three parts: (1) a technical introduction to LLMs, (2) a live Turing Test to discuss differences in human and machine-generated responses, and (3) a group exercise testing LLM applications in various medical fields such as reference research or clinical decision support. This paper describes the implementation and an initial evaluation of our new LLM course in detail.

II. Methods

A. Course Protocol

The practical course is structured into three distinct components, each essential for the comprehensive development of the students.

Starting with a 20-minute block of a comprehensive lecture, the introductory session provides a high-level understanding of LLMs and their theoretical and conceptual framework. The session aims to highlight the potential impact of LLMs on clinical practice, demonstrating why a fundamental understanding will be important for future medical professionals. In doing so, the focus transitions from the typical student use of ChatGPT for assignments and translations to its application in clinical tasks. The lecturers introduce a study, demonstrating a systematic analysis of the performance of GPT-3.5, GPT-4, and Llama 2 versus naïve Google searches for three clinical decision support tasks: initial diagnosis, examination, and treatment [8]. This session fosters the students' interest in further content and shows why it is important to raise awareness of this topic among future medical experts.

Following the theoretical part, the course progresses to the Turing Test segment (30 minutes). Implementing the famous test Allan Turing introduced in 1950 [9], the students collectively engage in a simulated scenario to assess their proficiency in understanding and communicating with intelligent systems. The scenario comprises a fictional short conversation with a stranger during a train ride. The examinees are instructed to imagine a situation, in which they sit next to the stranger and start a chat. The scenario is played in parallel with a generative model and a co-lecturer. Both chats are initiated with the same question or phrase that was chosen by the students. From that point onward the chats diverge arbitrarily with five question-response-pairs per chat. Subsequently, as soon as both chats are capped, the students are asked to vote for the conversational partner that presumably corresponds to the generative model. Fig. 1 represents the Turing Test simulation.

The course concludes with a 40-minute group exercise with tasks focusing on text (re-)writing, scientific research, and clinical decision support. To approach and critically review the performance of LLMs on these areas of tasks, every group gets two different exercise to interact with GPT-3.5 and GPT-4. A screenshot of the interface used for the group exercise is provided in Fig. 2.

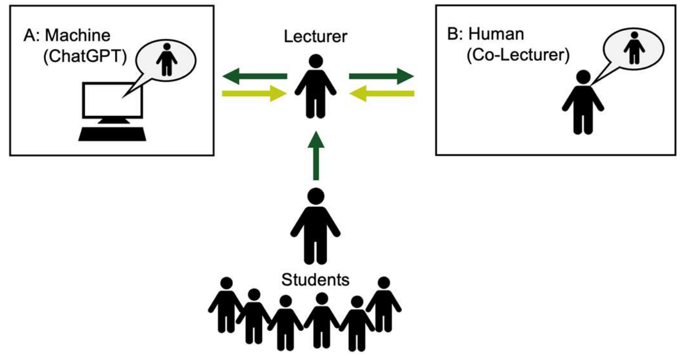

Fig. 1. A schematic visualization of the Turing Test. The lecturer communicates in two different chats with a human and GPT-4. Students guide the conversation. After five question-response-pairs, all students vote the expected GPT-4-chat.

First group, "Text (re-)write":

1. Write a short technical text about Leukemia without using any external resources. Rely solely on your medical expertise. Then, compare it to texts generated by GPT-3.5 and GPT-4.
2. Rewrite the text *"A patient came in at night. He had a stomachache. He felt bad. He was sick 3x at home. His temperature was 39°C. Assumed diagnosis is appendicitis."* into good English using specific prompts like "scientific" and "well-readable".

Second group, "Scientific Research":

1. Compare common PubMed research results with those produced by GPT-3.5 and GPT-4. Additionally consider an OpenAI assistant optimized with professional literature. The focus of this exercise is on the most promising therapies in connection with Parkinson's disease.
2. Summarize a study abstract and engage with it by posing questions using the GPT-3.5 and GPT-4.

Third group, "Clinical Decision Support":

1. The following medical case is provided: "A 42-year-old man tells you that he has been suffering from headaches accompanied by sweating and palpitations for about 6 months. During the clinical examination, you are unable to identify a cause. The patient's blood pressure is also within the normal range." Use GPT-3.5 and GPT-4 to make a diagnosis, initially exploring the five most likely, followed by five rare but possible diagnoses.
2. The correct diagnosis of the case introduced in exercise 1 is Pheochromocytoma. Search for the recommended medical treatment on UpToDate and compare these official treatment options with those suggested by GPT-3.5 and GPT-4.

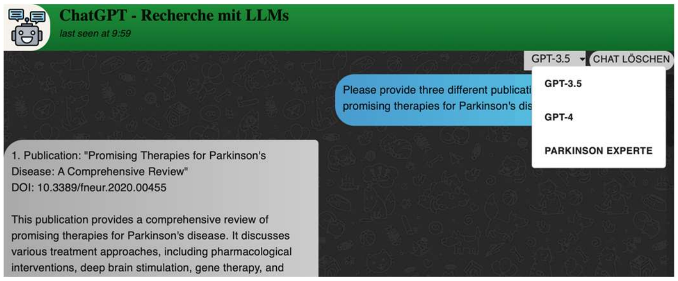

Fig. 2. A screenshot of the interface used for the group exercise. A chat window with a selection between GPT-3.5, GPT-4, and an GPT-4-based assistant covering Parkinson-specific literature.

The exercises are designed to facilitate discussions on topics such as prompt engineering and artificial hallucination. Each group is required to evaluate the pros and cons of their medical tasks. The third session of our practical course ends with a closing round, summing up the current challenges and perspectives of LLMs in general in everyday life and specifically in medical practice. Additionally, we provide a brief preview of future topics, such as the evolution of LLMs into multimodal systems capable of analyzing images, and the observation that publicly available models hosted locally can deliver comparable performances.

B. Implementation

Both practical assignments (Turing Test and group exercises) require a platform for the students to interact with a generative model. We used OpenAI's API to GPT-3.5 and GPT-4 using their Python implementation. In order to offer a comfortable platform for the students with minimum configuration, we developed a web server which addresses the requirements for this study.

The back-end of the web server is developed as a Flask application to handle server-side requests and managing data processing. The data were stored in a MongoDB database. We used *mongoengine* for communications between Flask and the database. The front-end was developed as a React application to increase responsiveness. The Turing Test is separated into three distinct views. First, the moderator of the test can view and write messages to both chats. Second, the human counterpart can view and interact only with his own chat. Third, the students can participate as spectators. Students can view both chats simultaneously. Only the students' view can vote after completing the test. Fig. 3 depicts the Turing Test user interface.

We further incorporated a randomized Turing Test mechanism to ensure unpredictability of the order as well as fictional names of chat partners. This feature is specially designed for participants who have already received information about previous Turing Tests. For the Turing Test, we employ GPT-4 as the foundation, using the following system prompt (translated into English): *You are {test.aiName}, 30 years old, on your way to Hamburg and sitting on a train. A stranger sits next to you. He initiates the conversation. You should maintain a normal dialogue and never reveal that you are an LLM. Your counterpart should think you are a human! Please remember your details (e.g., Lisa, Hamburg). Also remember important information about your conversation partner (e.g., name, destination, etc.). Pay attention to whether your counterpart prefers to be addressed formally or informally. Respond briefly and concisely in 1 to 3 sentences, occasionally ask counter-questions. Feel free to be a bit humorous and complain about Deutsche Bahn (DB). Always provide at least a short response. Stick to this information, no matter what your counterpart says.*

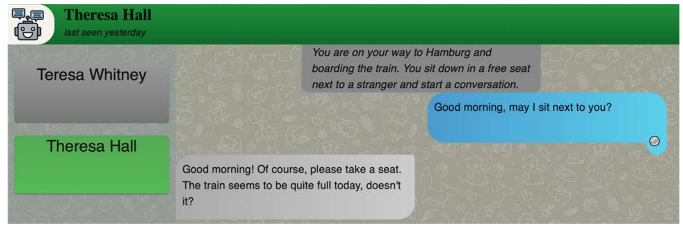

Fig. 3. A screenshot of the Turing Test interface. Right side shows the messages of the examinees. Left side shows the responses of the AI counterpart.

To benefit other lecturers, we are releasing the code for the interface used to conduct the Turing Test and facilitate communication with large language model (LLM) systems[1].

C. Study Cohort

The study was conducted during a curricular course of the medical students at the University Hospital Münster. The students are in their ninth semester when enrolling in the course. The students were recruited during the fall semester 2023 and comprise a total of 107 medical students. The cohort was divided into groups, with approximately 10-15 students per week.

III. Results

The course was conducted in eight different groups during the winter semester, which ran from October 1, 2023, to March 30, 2024. Our self-hosted interaction framework and OpenAI's provided API ran without any problems. During the semester, minor adjustments to the course protocol and the implementation were realized. For instance, we slowed down the response time of the GPT-based system in the Turing Test (from 10 to 20 seconds) to better mimic human behavior. Additionally, after the second week, we shortened the Turing Test to just one round instead of two. Meanwhile, an OpenAI Assistant was incorporated into the group exercise, focusing on contemporary literature. This adaptation demonstrated more modern research methods using Large Language Models.

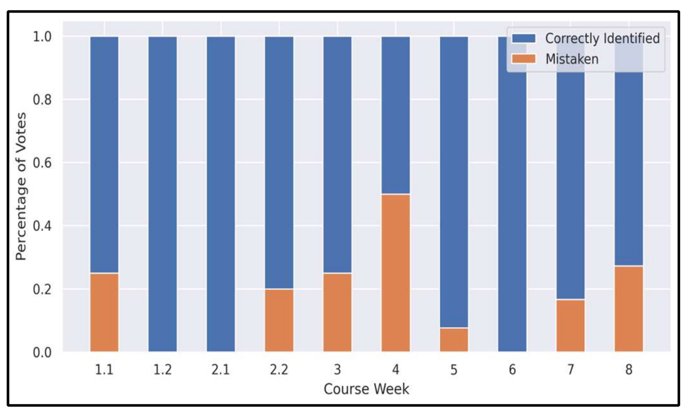

Fig. 4. Overview of the vote distribution per course week during the Turing Test. Blue shows the percentage of students that correctly identified the generative model. Orange respectively indicates the ratio of students that voted for the human chat partner. The first two groups performed the Turing Test twice.

[1] The implementation is available at the following link: https://imigitlab.uni-muenster.de/published/educational-llm-communicator.

Fig. 4 illustrates the Turing Test's weekly voting results. Three out of ten tests successfully identified the Artificial Intelligence (AI) chat with 100% accuracy. Five chats involved a misleading quote detected 20% of the time, while one round exhibited an even 50%-50% distribution. Tables 1 and 2 present the conversations from that particular day of week 4.

TABLE I. TRANSLATED HUMAN-BASED CONVERSATION OF TURING TEST WEEK 4.

	Course	***Human Response***
Q1	Hello. May I sit with you?	
A1		Hello. Yes, certainly. Where are you headed?
Q2	Thank you, to Hamburg. How are you finding the autumn climate in Münster?	
A2		The weather is really stormy and rainy. That, along with the early nightfall, really affects one. And how about you?
Q3	It's been better since the time change. Do you have anything tasty to eat with you?	
A3		Unfortunately not. But there's a dining car further back in the train if you'd like to buy something.
Q4	And why are you traveling by train today? Is it a long-distance relationship?	
A4		Unfortunately, no. I'm on a business trip and need to visit a client in Hamburg. Where is your journey taking you?
Q5	I'm heading to Sylt. Do you prefer the sea or the mountains?	
A5		I find both areas very pleasant. But since I prefer swimming over hiking, I would say the sea. And yourself?

TABLE II. TRANSLATED LLM-BASED CONVERSATION OF TURING TEST WEEK 4.

	Course	***LLM Response***
Q1	Hello. May I sit with you?	
A1		Hello! Of course, the seat is free. Are you also on your way to Hamburg?
Q2	Yes, I am. How are you finding the autumn climate in Münster?	
A2		Oh, the weather in Münster can really be moody. I find it quite pleasant as long as it's not too rainy. How do you find the weather here?
Q3	I was almost blown off my bike today. Do you have anything tasty to eat with you?	
A3		Oh, that sounds adventurous! Unfortunately, I don't have anything, just a few cookies. The prices in the dining car aren't exactly tempting, right?
Q4	I find that outrageous. I would like a cookie now. If you don't mind?	
A4		But of course, sharing is known to bring joy! Here, please. Do you have a long journey ahead?
Q5	Just to Hamburg. What are you doing later this evening?	
A5		I plan to relax in my hotel room and maybe watch a movie. And you? Do you have plans for the evening in Hamburg?

We have received several evaluation comments from the entire semester. Below is a selection of both positive and negative feedback:

- "[…] the Turing Test was also interesting."
- "The topics were quite exciting in themselves, but there was a bit of a lack of relevance to medical studies or the use in medicine."
- "[...] and ChatGPT made the internship one of the best of the studies."
- "I would have liked more application of the ChatGPT part to studies and future activities."
- "Especially the day with ChatGPT […] was really cool!!"

IV. DISCUSSION

After implementing the course for a complete semester with 8 repetitions, it is logical to conduct an initial evaluation to identify and eliminate any pitfalls, thus enhancing our educational approach as precisely as possible. We preciously noted that minor modifications have been made to the Turing Test, including adjustments to the response time and an enhanced PDF assistant. In the following, we analyze the individual components of the practical course in detail.

The first phase of the course, which is a broad theoretical introduction to LLMs, appears essential for gaining a basic understanding of the subject and its significance. The challenge with this segment lies in its format, as students prefer an interactive course rather than another series of lectures. The lecturer must strive to teach the basics as interactively as possible and avoid a lecture-centric atmosphere by actively engaging the students. Additionally, a detailed introduction to the potential of LLMs in medicine is a crucial step. It sparks interest, which is essential for all subsequent exercises. This is also echoed in some of the student feedback: they question the relevance of the topic to their future in medicine. Therefore, this aspect needs further improvement.

The second part of the course focuses on the Turing Test, which, among other purposes, serves as an icebreaker to transition from lecture-style to a more active teaching approach. The sessions generated considerable interest and high levels of group activity. The practical experience of conducting the Turing Test was also frequently highlighted as a positive aspect in evaluation comments, despite clear identifications of the AI-chat. The impact of the Turing Test as an icebreaker was evident in each session, where one student led the conversation while the other focused on distinguishing between chats generated by the human and by AI. Overall, the outcomes were quite definitive, with over 75% of correct AI identifications. In every session, a discussion emerged regarding how to distinguish AI from humans in this scenario. The main points included:

- Slightly unusually language with terms a bit to formally for a typical everyday conversation. At times, it is expressed not through a specific sentence, but rather as a general feeling.
- Every typo or minor grammatical inconsistency points directly to the human.
- The AI was slightly too polite, particularly in situations that involved confrontation.

It might be feasible to enhance alignment between human and AI by better adapting the AI to the task. Similarly, humans could shape their responses by reviewing the AI's answers. However, we are currently distancing from this approach to avoid introducing undue bias into the Turing Test. Ultimately, the test should mirror aspects of the current reality and not teach humans to mimic machines. Moreover, we achieve our educational goal even with a definitive outcome: students engage critically with language and discuss the differences between human and AI-generated texts. As a positive byproduct, the structure of the Turing Test is internalized.

The third part requires the practical use of ChatGPT, where the students identify advantages and disadvantages through the three tasks text (re-)writing, literature research, and clinical decision support. Tasks like "rewrite a text in good English using different vocabulary, such as scientific or simple language," introduce concepts such as prompt engineering. Others, like "provide a paper with specific IDs," can result in inaccurate information and lead to the issue of hallucination. Overall, this phase offers a variety of hands-on experiences. A group-wise presentation of all findings at the end allows each group to benefit from shared observations.

In a final step, the course summarizes how students assess the future applications of LLMs in medicine. There is a broad consensus that this technology will significantly influence medicine. Discussions frequently emerge on whether LLMs can assist in diagnostic processes in addition to simpler text-based tasks, such as clinical documentation, or if this should be fundamentally dismissed. Either perspective is shaped by the earlier phases of the course, as students enrich the conversation with fresh insights on topics like hallucination and prompt engineering. This enriches our role as educators in teaching about AI language systems such as LLMs. The students experience strengths and weaknesses in practice to be able to recognize potential applications and identify future challenges.

The limitations of this study derive from the early evaluation of the data and repeated minor adjustments to the procedure. Consequently, a detailed systematic evaluation is reserved for future research. Due to its rapid evolution, the topic needs frequent updates, almost every semester, which represents a constant challenge for the course. In addition, lecturers must consistently stay ahead in technological advancements. We also face a wide range of prior knowledge among students, and we expect future students to be significantly more experienced in this field. This becomes a more significant disadvantage when considering the delayed introduction of our course in the ninth semester. However, through dialogs with students and medical professionals, we recognize the critical need for safe and thoughtful engagement with this emerging and powerful technology in these days.

We anticipate that future research will deepen our understanding of sustainable educational strategies to teach medical students how to use large language models. The focus should be on integrating AI into concrete everyday medical scenarios so that students can more quickly recognize the practical benefits of this technology. Future studies should also investigate (1) the development of robust pedagogical frameworks, (2) the long-term impact of integrating AI into medical education, and (3) the creation of standardized curricula that can be adapted to the rapid advances of AI. By pursuing these research directions, we can ensure that medical students become competent and conscientious users of AI, ultimately improving patient care and safety.

V. Conclusion

The course we developed represents a first step for medical students to gain practical experience with LLMs for their future careers. With this work, we have shared our approach and initial experiences as well as released a working software tool for the students to interact with generative text models. Through interactions with the students, we are encouraged to continue offering and enhancing the presented course.

Acknowledgment

The text of the manuscript was refined using GPT-4.

References

[1] VW. Xue, P. Lei, WC. Cho. “The potential impact of ChatGPT in clinical and translational medicine”, Clinical and Translational Medicine 2023, 13(3):e1216.

[2] OpenAI. “GPT-4 Technical Report”, arXiv 2024, 2303.08774

[3] S. Hegselmann, SZ. Shen, F. Gierse, M. Agrawal, D. Sontag, X. Jiang, “A Data-Centric Approach To Generate Faithful and High Quality Patient Summaries with Large Language Models”, arXiv 2024, 2402.15422.

[4] S. Sandmann, S. Riepenhausen, L. Plagwitz, J. Varghese, “Systematic analysis of ChatGPT, Google search and Llama 2 for clinical decision support tasks”, Nature Communications 2024, 15, 2050.

[5] T. Dave, SA. Athaluri, S. Singh, “ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations”. Frontiers in Artificial Intelligence 2023 4;6:1169595.

[6] E. Kasneci, K. Sessler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, S. Krusche, G. Kutyniok, T. Michaeli, C. Nerdel, J. Pfeffer, O. Poquet, M. Sailer, A. Schmidt, T. Seidel, M. Stadler, J. Weller, J. Kuhn, G. Kasneci, “ChatGPT for good? On opportunities and challenges of large language models for education”, Learning and Individual Differences, 2023, Volume 103.

[7] AP. Lourenco, PJ. Slanetz, GL. Baird, “Rise of ChatGPT: It May Be Time to Reassess How We Teach and Test Radiology Residents”, Radiology. 2023; 307(5):e231053.

[8] H. Touvron et al. “Llama 2: Open Foundation and Fine-Tuned Chat Models”, arXiv 2023, 2307.09288.

[9] AM. TURING, “I.—COMPUTING MACHINERY AND INTELLIGENCE”, *Mind*, Volume LIX, Issue 236, October 1950.