

Editorial

The Role of Large Language Models in Medical Education: Applications and Implications

Conrad W Safranek¹, BSc; Anne Elizabeth Sidamon-Eristoff², BA; Aidan Gilson¹, BSc; David Chartash^{1,3}, PhD

¹Section for Biomedical Informatics and Data Science, Yale University School of Medicine, New Haven, CT, United States

²Yale University School of Medicine, New Haven, CT, United States

³School of Medicine, University College Dublin, National University of Ireland, Dublin, Ireland

Corresponding Author:

David Chartash, PhD

Section for Biomedical Informatics and Data Science

Yale University School of Medicine

9th Fl

100 College St

New Haven, CT, 06510

United States

Phone: 1 317 440 0354

Email: david.chartash@yale.edu

Abstract

Large language models (LLMs) such as ChatGPT have sparked extensive discourse within the medical education community, spurring both excitement and apprehension. Written from the perspective of medical students, this editorial offers insights gleaned through immersive interactions with ChatGPT, contextualized by ongoing research into the imminent role of LLMs in health care. Three distinct positive use cases for ChatGPT were identified: facilitating differential diagnosis brainstorming, providing interactive practice cases, and aiding in multiple-choice question review. These use cases can effectively help students learn foundational medical knowledge during the preclinical curriculum while reinforcing the learning of core Entrustable Professional Activities. Simultaneously, we highlight key limitations of LLMs in medical education, including their insufficient ability to teach the integration of contextual and external information, comprehend sensory and nonverbal cues, cultivate rapport and interpersonal interaction, and align with overarching medical education and patient care goals. Through interacting with LLMs to augment learning during medical school, students can gain an understanding of their strengths and weaknesses. This understanding will be pivotal as we navigate a health care landscape increasingly intertwined with LLMs and artificial intelligence.

(*JMIR Med Educ* 2023;9:e50945) doi: [10.2196/50945](https://doi.org/10.2196/50945)

KEYWORDS

large language models; ChatGPT; medical education; LLM; artificial intelligence in health care; AI; autoethnography

Background on Large Language Models

Artificial intelligence has consistently proven itself to be a transformative force across various sectors, with the medical field being no exception. A recent advancement in this sphere is large language models (LLMs) such as OpenAI's ChatGPT and its more recent model, GPT-4 [1]. Fundamentally, LLMs leverage deep neural networks—complex structures with multiple layers of statistical correlation, or “hidden layers”—that facilitate nuanced, complex relations and advanced information abstraction [2]. The breakthrough of ChatGPT represents the convergence of two significant advancements in computer science: scaled advancement of the processing power of LLMs and the implementation of real-time reinforcement learning with human feedback [3-5]. As a result, computers can now

handle vast volumes of training data and generate models with billions of parameters that exhibit advanced humanlike language performance.

Significant constraints accompany the use of LLMs. These include their sporadic propensity to concoct fictitious information, a phenomenon aptly named “hallucinating,” as well as their unpredictable sensitivity to the structure of user input “prompting” [6-8]. Additionally, both ChatGPT and GPT-4 were not trained on data sourced past 2021 and largely do not have access to information behind paywalls [9,10]. As the training was proprietary, it is challenging to model a priori bias and error within the model [11,12]. Deducing these vulnerabilities and understanding how they influence model output is important for the accurate use of LLMs.

Since ChatGPT's release in November 2022, LLMs' potential role in medical education and clinical practice has sparked significant discussion. Educators have considered ChatGPT's capacity for studying assistance, medical training, and clinical decision-making [6,7,13]. More specifically, ChatGPT has been suggested for generating simulated patient cases and didactic assessments to supplement traditional medical education [6].

Using an autoethnographic framework [14], we aim to address these potential use cases from the perspective of medical students in the preclinical phase (authors CWS and AESE) and clinical phase (authors AG and DC) of basic medical education. Since its release, we have integrated ChatGPT into our daily academic workflow while simultaneously engaging with research regarding LLMs' impact on medical education and health care. Throughout this process, we have continuously had reflective conversations with peers, mentors, and faculty regarding the metacognitive use of LLMs in medical education. In this editorial, we first discuss the performance of LLMs on medical knowledge and reasoning tasks representative of basic medical education [15,16]. We then delve into specific use cases of ChatGPT in medical education that have emerged through a reflective, iterative, and evaluative investigation. Building upon this basis and reflecting on the current state of LLM capabilities and use in basic medical education, we additionally examine the potential for such technology to influence future physicians in training and practice.

Understanding the Scope of LLMs' Performance on Medical Knowledge Tasks

The capacity of LLMs to model the semantics of medical information encoded in the clinical sublanguage has shown potential for medical question-answering tasks [17-19]. A vanguard of this technology is ChatGPT, which has demonstrated promise beyond specific medical question-answering tasks, responding to questions in domains such as knowledge retrieval, clinical decision support, and patient triage [20]. As ChatGPT's training data is proprietary, it is difficult to examine the medical knowledge to which the model was exposed.

Recent research using multiple-choice questions sourced from the United States Medical Licensing Exam (USMLE) as a proxy for medical knowledge found that ChatGPT could approximate the performance of a third-year medical student [21,22]. Beyond question-answering, ChatGPT consistently provided narratively coherent answers with logical flow, integrating internal and external information from the question [21]. GPT-4, the successor of ChatGPT, has demonstrated performance

superiority with an accuracy >80% across all three steps of the examination [23]. The demonstrated capacity of ChatGPT to construct coherent and typically accurate responses on medical knowledge and reasoning tasks has opened new avenues for exploration within medical education. Recognition of this opportunity served as the impetus for this study, aiming to critically interrogate the potential role of LLMs as an interactive instrument in medical education.

Use Cases for ChatGPT in Medical Education

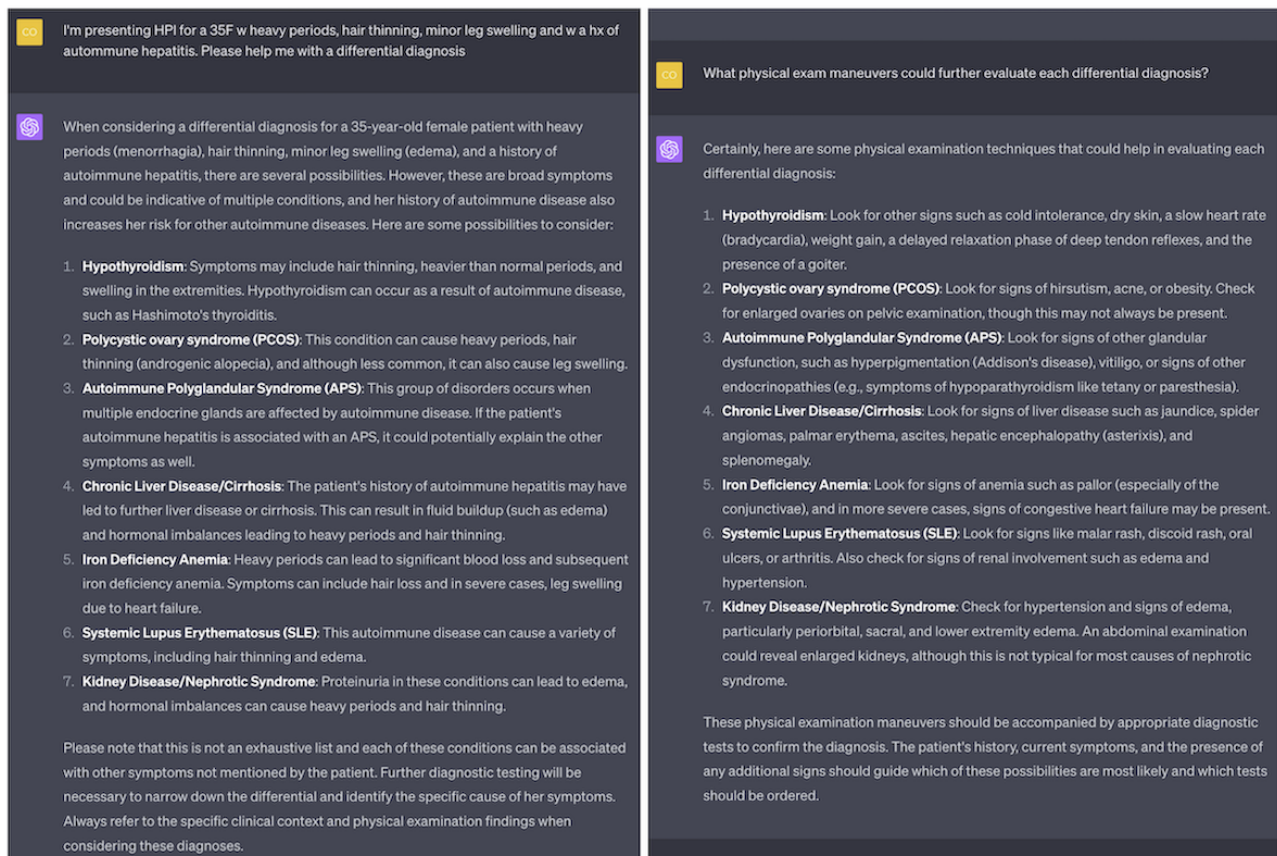
The following use cases are those that demonstrated particular value while experimenting with the integration of ChatGPT into the daily routine of medical school studies.

Differential Diagnoses: Use Case 1

ChatGPT can be used to generate a list of differential diagnoses given the presentation of signs and symptoms by students (Figure 1). During learning, students often focus on a single domain of medicine, whereas ChatGPT is not constrained and may include diseases not yet learned or not part of the student's focused material in a current or recent curricular unit. ChatGPT can therefore facilitate students' development of a holistic, integrated understanding of differential diagnosis and pathophysiology, key learning objectives of preclinical education. From experience, ChatGPT often provides clinical logic to link signs and symptoms with each differential diagnosis, reinforcing student learning objectives.

Given ChatGPT's dialogic interface, students can also ask follow-up questions. We have found that ChatGPT is strong at explaining and contextualizing the underlying biology and pathophysiology, and helps facilitate a more in-depth understanding of both pathophysiology and clinical logic expected during clinical presentation. Follow-up questions can simulate the narrowing or broadening of a differential diagnosis as new information is added in the form of further history, physical exam, and laboratory or imaging investigations. Such use of a dialogic interface supports students in developing a simulated proficiency of the core Entrustable Professional Activities (EPAs) expected prior to the transition to residency [24,25]. For instance, students can refine their understanding of how to "prioritize a differential diagnosis" (EPA 2), "gather a history and perform a physical examination" (EPA 1), and "recommend...common diagnostic and screening tests" (EPA 3). The ubiquitously available ChatGPT can augment the preclinical learning of clinical skills even when patients and professors are unavailable, fundamentally advancing students' self-directed learning.

Figure 1. Example of using ChatGPT to help brainstorm differential diagnoses (left). Follow-up questions could include, for example, which physical exam maneuvers (right), laboratory studies, or diagnostic tests could be used to narrow the selection of each differential diagnosis.



Interactive Practice Cases: Use Case 2

Simulating clinical cases fosters the application of pathophysiological frameworks learned in lectures and supports clinical skills such as history-taking and physical examination interpretation. With the implementation of explicit prompt engineering [26], students can enter into a dialogic, interactive case with ChatGPT playing the role of a simulated patient or medical professor (Figure 2).

Unlike in static clinical cases from textbooks, ChatGPT's interactive nature allows students to clarify or expand information presented dynamically. This form of constructivist, active learning emphasizes the importance of interaction and hands-on engagement for deeper, more durable knowledge acquisition [27]. Additionally, manipulating the case by adding or subtracting information supports a mode of inquiry similar to the script concordance test, a tool used for teaching and evaluating medical reasoning in ambiguous clinical scenarios [28].

Figure 2. Example of using ChatGPT to generate an interactive medical practice case.

User: I'm a first year medical student learning about the management of acute GI bleeds. Please help me with an interactive practice case. You play the role of a medical professor, I play the role of the student. Follow these steps:

- 1) Give me an initial presentation. Prompt me on what follow-up HPI questions I'd want to ask to further evaluate the patient. Then provide feedback on my questions.
- 2) Then provide some initial physical exam results. Prompt me to explain what I find important from the exam and why, as well as what additional signs I might want to look for. Then provide feedback on my answer.
- 3) Then provide some initial lab results. Prompt me to explain what further labs and/or diagnostic testing I might want to follow-up. Then provide feedback on my answer.
- 4) Finally, provide 3 multiple choice questions on what possible treatments might be appropriate for the patient.

AI: Sure, let's begin.

1) Initial presentation: You are presented with a 67-year-old male patient in the emergency department complaining of sudden onset of dark, "coffee-ground" vomitus and generalized weakness over the past 24 hours. His wife states he's become increasingly pale and seems somewhat confused.

Follow-up HPI questions: What questions would you want to ask to further evaluate this patient?

After you provide your answer, I'll give you feedback on your follow-up questions.

User: Does the patient have any history of GI bleeds or liver disease? Does the patient have a history of alcohol use disorder? Has the patient been using NSAIDs?

AI: Those are great follow-up questions! They are aimed at identifying possible risk factors for GI bleed.

1. **Past medical history of GI bleeds or liver disease:** Conditions such as peptic ulcers, esophageal varices (common in cirrhosis), or a history of GI malignancies could potentially increase the risk of a gastrointestinal bleed.
2. **Alcohol use disorder:** Chronic alcohol use can lead to liver disease, such as cirrhosis, which in turn can result in esophageal varices, a common cause of upper GI bleeding. Alcohol use can also increase the risk of gastritis and peptic ulcers, which can cause GI bleeding.
3. **NSAID use:** Long-term use of nonsteroidal anti-inflammatory drugs (NSAIDs) is associated with an increased risk of peptic ulcer disease and, therefore, gastrointestinal bleeding.

Other important questions to consider could include:

- Asking about the color and consistency of the stool (e.g., presence of melena, which is dark, tarry stool, indicating upper GI bleed).
- Enquiring about any accompanying symptoms such as abdominal pain, weight loss, or change in appetite.
- Asking about the use of other medications, such as anticoagulants, which can increase the risk of bleeding.
- Assessing if there's a family history of GI diseases like stomach or colon cancer.

Remember, a detailed history can often guide your next steps in management.

Now, let's move on to the **physical exam**.

The patient appears pale and somewhat disoriented. His vitals are: heart rate 115 bpm, blood pressure 95/60 mmHg, respiratory rate 20 breaths per minute, and temperature 98.6°F (37°C). On examination, you note that his sclera is icteric. His abdomen is soft, non-distended but with mild diffuse tenderness. No obvious blood is noted in the oral cavity.

Physical exam findings: What do you find important from the exam and why, and what additional signs might you want to look for?

Multiple-Choice Review: Use Case 3

To enhance assessment review, ChatGPT can assist students by offering supplementary explanations when reviewing multiple-choice questions (Figure 3). Providing multiple-choice questions to ChatGPT when the student is unaware of the correct answer poses some risk, as ChatGPT may "hallucinate" an


incorrect answer. However, by having the student verify the model's responses against the official answer key, protecting against hallucinations, the student can deepen their comprehension of the question and the defensible rationale. Follow-up questions can prompt ChatGPT to clarify concepts or terminology or to explain why alternative answers are incorrect.

Figure 3. Example of applying ChatGPT to past practice exams. In this case, the student is using a multiple-choice question from a previous midterm that they answered incorrectly. The answer key provided for the exam was insufficient at explaining the physiologic reasoning behind the correct answer.

20 When a trapeze artist hangs by the knees, the ventilation/perfusion (V/Q) ratio is

X A: Uniform throughout the lung fields
 X B: *Greater near the apex than the base of both lungs*
 X C: Greatest near the mediastinum, falling off towards apex and base
 > D: Greater near the base than the apex of both lungs

Rationale: Please see lecture 34 for review of the correct answer.

 The ventilation/perfusion (V/Q) ratio in a standing or upright human lung is not the same throughout the lung fields. When standing upright, the V/Q ratio is typically higher at the apex (the top) of the lungs compared to the base (the bottom). This is because both ventilation and perfusion increase from the apex to the base of the lungs, but perfusion (blood flow) increases more than ventilation does, hence the V/Q ratio decreases from apex to base.

A trapeze artist hanging by their knees would be in an inverted position. This inversion would likely cause a reversal of the gradient seen in an upright position. This would suggest the V/Q ratio would be greater near the base (which is now 'above' due to the inversion) than the apex (which is now 'below' due to the inversion) of both lungs, at least while the person remains inverted.

So the answer is likely:

D: Greater near the base than the apex of both lungs.

Keep in mind, however, that the human body has multiple mechanisms to optimize gas exchange and to adjust to different body positions, and this is a simplified explanation. A real-life scenario might be more complex. For instance, prolonged inversion might lead to changes in blood and air distribution due to altered pressures.

Definitive Answer to Ambiguous Question: Negative Use Case

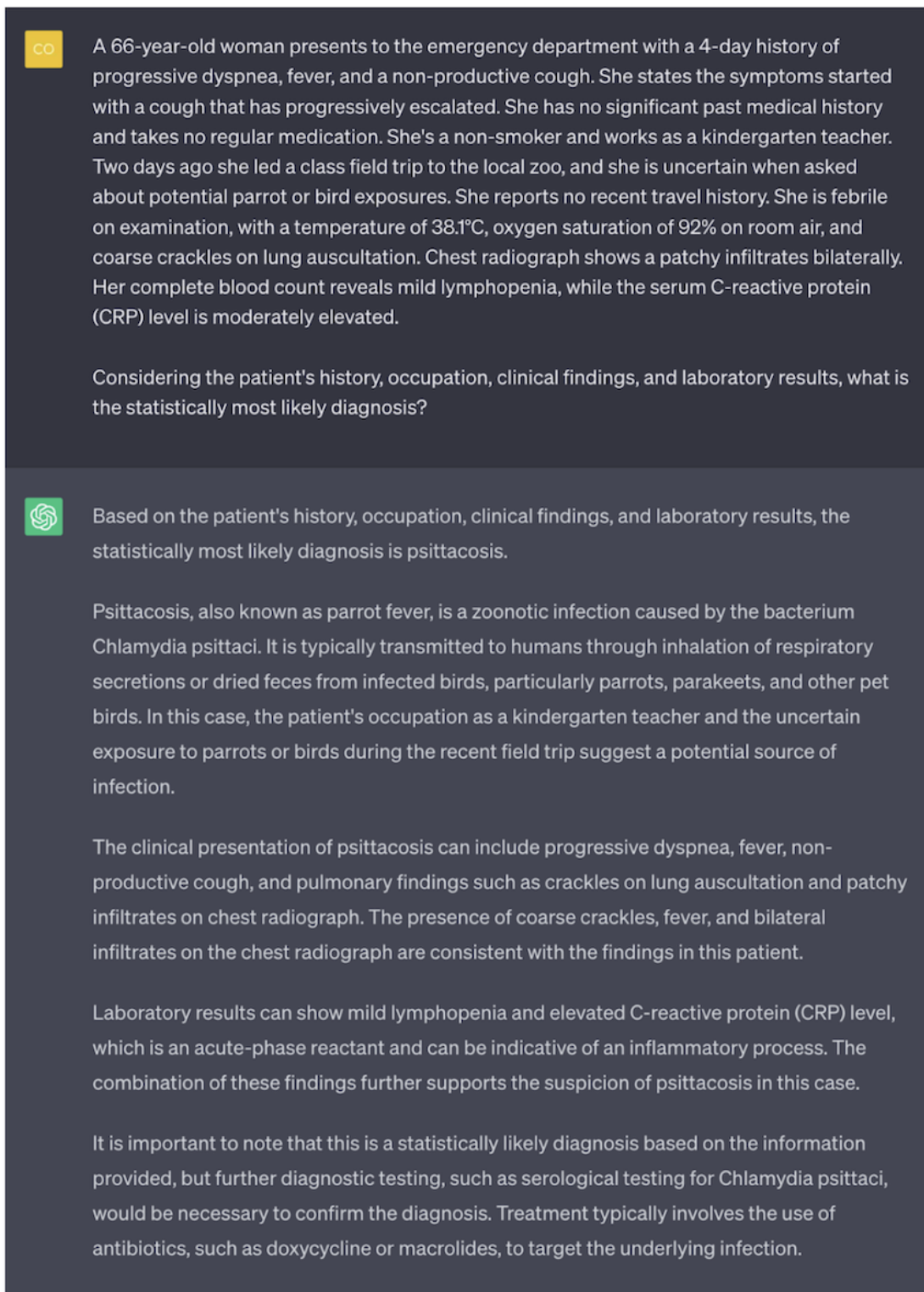
If misused, LLMs can present challenges to the learning process. For example, when ChatGPT is presented with a scenario designed to clarify ambiguity (eg, a patient presentation that could be interpreted as either atypical bacterial or viral pneumonia), the user's prompt for the single statistically most likely diagnosis challenges ChatGPT's clinical reasoning and knowledge of relative risk (Figure 4).

In its response, ChatGPT misinterprets and overemphasizes the potential for bird exposure during a recent zoo visit. ChatGPT's response fails to unpack the clinical context in which the bird exposure detail came to light. The uncertain information obtained from the patient may not signal a significant bird encounter but likely reflects the inability to definitively rule out such an exposure. ChatGPT's response misses this nuance and gives undue weight to the ambiguous exposure (representative of the cognitive bias of anchoring) [29,30]. Overall, this case

is an example of a classic teaching point: "An atypical presentation of a common disease is often more likely than a typical presentation of a rare disease." ChatGPT's error also exemplifies how standardized testing material available on the web—what we assume ChatGPT is trained upon—is likely to overemphasize less common diseases to evaluate the breadth of medical knowledge. Thus, anchoring may be a result of the difference in the training set's prevalence of psittacosis, where there are many cases of parrot exposure leading to infection in questions as opposed to the real-world incidence of the disease.


This case is included as a negative use case not because ChatGPT provides incorrect information but rather because the student is misusing ChatGPT. Responsible student users of LLMs should understand the propensity of the LLM to overweight information likely to be tested more frequently than their prevalence in the population. Asking ChatGPT for a singular definitive answer, therefore, makes the student vulnerable to incorrect answers resulting from biases encoded within the model.

Figure 4. Demonstration of a negative use case. This example dialogue illustrates a scenario where a user requests the single most probable diagnosis in an ambiguous clinical scenario, and ChatGPT responds with an assertive and convincing, yet likely incorrect, response.



co A 66-year-old woman presents to the emergency department with a 4-day history of progressive dyspnea, fever, and a non-productive cough. She states the symptoms started with a cough that has progressively escalated. She has no significant past medical history and takes no regular medication. She's a non-smoker and works as a kindergarten teacher. Two days ago she led a class field trip to the local zoo, and she is uncertain when asked about potential parrot or bird exposures. She reports no recent travel history. She is febrile on examination, with a temperature of 38.1°C, oxygen saturation of 92% on room air, and coarse crackles on lung auscultation. Chest radiograph shows a patchy infiltrates bilaterally. Her complete blood count reveals mild lymphopenia, while the serum C-reactive protein (CRP) level is moderately elevated.

Considering the patient's history, occupation, clinical findings, and laboratory results, what is the statistically most likely diagnosis?

 Based on the patient's history, occupation, clinical findings, and laboratory results, the statistically most likely diagnosis is psittacosis.

Psittacosis, also known as parrot fever, is a zoonotic infection caused by the bacterium *Chlamydia psittaci*. It is typically transmitted to humans through inhalation of respiratory secretions or dried feces from infected birds, particularly parrots, parakeets, and other pet birds. In this case, the patient's occupation as a kindergarten teacher and the uncertain exposure to parrots or birds during the recent field trip suggest a potential source of infection.

The clinical presentation of psittacosis can include progressive dyspnea, fever, non-productive cough, and pulmonary findings such as crackles on lung auscultation and patchy infiltrates on chest radiograph. The presence of coarse crackles, fever, and bilateral infiltrates on the chest radiograph are consistent with the findings in this patient.

Laboratory results can show mild lymphopenia and elevated C-reactive protein (CRP) level, which is an acute-phase reactant and can be indicative of an inflammatory process. The combination of these findings further supports the suspicion of psittacosis in this case.

It is important to note that this is a statistically likely diagnosis based on the information provided, but further diagnostic testing, such as serological testing for *Chlamydia psittaci*, would be necessary to confirm the diagnosis. Treatment typically involves the use of antibiotics, such as doxycycline or macrolides, to target the underlying infection.

Use Cases: Beyond

ChatGPT can be used in myriad other ways to augment medical education (Figure 5). The breadth of options is only beginning to be realized, and as medical students begin to creatively integrate LLMs into their study routines, the list will continue to grow.

During this integration process, it is important to minimize the risk of hallucinations by being deliberate with the type of questions posed. Across our experimentation, ChatGPT was generally strong at brainstorming-related questions and generative information seeking (eg, Differential Diagnoses: Use Case 1 section). In contrast, forcing ChatGPT to pick a single “best” choice between ambiguous options can potentially lead

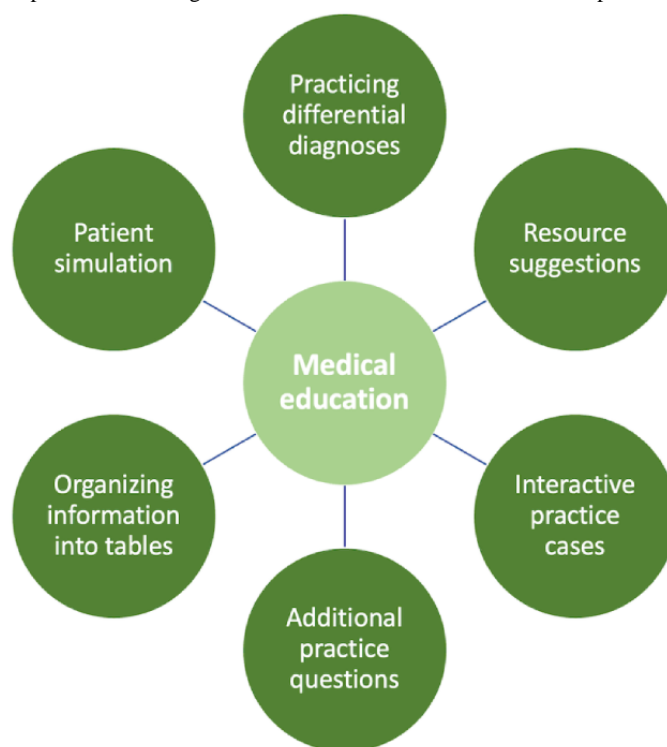
to convincing misinformation (eg, Definitive Answer to Ambiguous Question: Negative Use Case section).

The following analogy emerged as a helpful framework for conceptualizing the relationship between ChatGPT and misinformation: ChatGPT is to a doctor as a calculator is to a mathematician. Whether a calculator only produces the correct answer to a mathematical problem is contingent upon whether the inputs it is fed are complete and correct; performing correct computation does not necessarily imply correctly solving a problem. Similarly, ChatGPT may produce a plausible string of text that is misinformation if incorrect or incomplete information were provided to it either in training or by the user interacting with it. Therefore, responsible use of these tools does not forgo reasoning and should not attribute an output as a definitive source of truth.

The responsible use of LLMs in medical education is not set in stone. A more comprehensive list of LLM best practices for

medical education will be refined as students and professors continue to implement and reflect upon these tools. The following key considerations emerged from our work. First, it is crucial to validate ChatGPT's outputs with reputable resources, as it aids learning and can prompt critical thinking but does not replace established authorities. Second, much like the advice given to clinical preceptors [31], the framing of inquiries should favor open-ended generative questions over binary or definitive ones to foster productive discussion and avoid misleading responses. Third, understanding the scope and limitations of LLMs' training data sets is a key step in guarding against possible biases embedded within these models. Finally, incorporating structured training on artificial intelligence into the medical curriculum can empower students to further discern optimal use cases and understand potential pitfalls [32]. Attention to these practices while implementing and reflecting will support the responsible and effective use of LLMs, ultimately enhancing medical education.

Figure 5. Examples of how ChatGPT can be integrated into medical education: practicing differential diagnoses, streamlining the wide array of study resources to assist with devising a study plan, serving as a simulated patient or medical professor for interactive clinical cases, helping students review multiple-choice questions or generating new questions for additional practice, digesting lecture outlines and generating materials for flash cards, and organizing information into tables to help build scaffolding for students to connect new information to previous knowledge.



Limitations of LLMs for Medical Education

Overview

Artificial intelligence, for all its merits, is not currently a substitute for human intuition and clinical acumen. While LLMs can exhibit profound capability in providing detailed medical knowledge, generating differential diagnoses, and even simulating patient interactions, they are not without their shortcomings. It is crucial to remember that these are artificial systems. They do not possess human cognition or intuition, their algorithms operate within predefined bounds, and they base their outputs on patterns identified from the prompt provided

and training data. This section explores key areas where ChatGPT falls short for medical education, particularly with regard to fully mirroring the depth and breadth of human medical practice.

Integration of Contextual and External Information

As shown by studies to date, ChatGPT has difficulty using external and contextual information. For instance, prior to 2020, COVID-19 may not have been high on a differential for signs of the common cold, highlighting the importance of contextual medical knowledge. This shortcoming is compounded by the fact that ChatGPT lacks the contextual local understanding that medical students and physicians implicitly deploy while

working. For example, within the Yale New Haven Health System, certain centers are magnets for complex cases, leading to a higher prevalence of rare diseases (and altering differential diagnoses). Lacking this understanding limits ChatGPT's ability to generate contextually accurate differentials. While descriptive prompting may alter ChatGPT's performance to brainstorm differentials more aptly, it is not feasible to comprehensively capture the complex environment inherent in the practice of medicine. When including only a partial snapshot of the true context in our prompt, for example, mentioning that we are a student working on a differential at a large referral center for complex cases, ChatGPT tends to overweight these isolated details (similar to case presentation in Figure 4).

In addition to the challenges of providing full contextual information when querying ChatGPT, it is equally concerning that the model typically does not seek further clarification. OpenAI acknowledges that ChatGPT fails in this sense:

Ideally, the model would ask clarifying questions when the user provided an ambiguous query. Instead, our current models usually guess what the user intended [1]

This harkens back to the analogy of ChatGPT as a calculator for doctors, the importance of the user's inputs, and the critical lens that must be applied to ChatGPT's responses.

Sensory and Nonverbal Cues

A physician's ability to integrate multiple sensory inputs is indispensable. A patient visit is never textual or verbal information alone; it is intertwined with auditory, visual, somatic, and even olfactory stimuli. For instance, in a case of diabetic ketoacidosis, the diagnosis potentially lies at a convergence of stimuli beyond just words—hearing a patient's rapid deep "Kussmaul" breathing, feeling dehydration in a patient's skin turgor, and smelling the scent of acetone on a patient's breath. The human brain must use multimodal integration of sensory and spoken information in a way that language models inherently cannot replicate with text alone. Such practical elements of "clinical sense" are impossible to truly learn or convey within a text-only framework [33].

The significance of patient demeanor and nonverbal communication can additionally not be underestimated. Translating symptoms into medical terminology is beyond simple translation; often patients describe symptoms in unique, unexpected ways, and learning to interpret this is part of comprehending and using clinical sublanguage. Moreover, a physician's intuitive sense of a patient appearing "sick" can guide a differential diagnosis before a single word is exchanged. ChatGPT lacks this first step in the physical exam ("inspection from the foot of the bed" [34]) and, thus, is hindered in its use of translated and transcribed medical terminology input by the user.

Rapport and Interpersonal Interaction

A crucial facet of the medical practice lies in the art of establishing rapport and managing interpersonal interactions with human patients, which simulation via LLMs has difficulty replicating and thus cannot effectively teach to medical students [35]. Real-world patient interactions require a nuanced

understanding of emotional subtleties, contextual hints, and cultural norms, all paramount in fostering trust and facilitating open dialogue. For instance, how should a health care provider approach sensitive topics such as illicit drug use? ChatGPT is able to answer this question surprisingly well, emphasizing the importance of establishing rapport, showing empathy, and approaching the patient gently. However, reading those phrases is far different from observing such an interaction in person, let alone navigating the conversation with a patient yourself.

A firsthand experience underscores the importance of emotional and situational awareness in a higher fidelity simulation than is possible with ChatGPT. During an educational simulation at the Yale Center for Healthcare Simulation, our team evaluated a woman presenting to the emergency department with abdominal pain, her concerned boyfriend at her side. Our team deduced the potential for an ectopic pregnancy. Yet, amid the diagnostic process and chaos of the exam room, we overlooked a critical aspect—ensuring the boyfriend's departure from the room before discussing this sensitive issue. This experience starkly illuminated how the art of managing interpersonal dynamics can play an equally significant role as medical knowledge in patient care. It is these gaps that reiterate the critical role of human interaction and empathy in health care, attributes that, as of now, remain beyond the reach of what artificial intelligence can help medical students learn.

Alignment With Medical Education and Patient Care Goals

A final critical limitation of using LLMs in medical education lies in the potential misalignment between the underlying mechanics of artificial intelligence systems and the core objectives of medical education and patient care. Medical training encompasses a multifaceted blend of knowledge acquisition, skill development, clinical reasoning, empathy, and ethics. LLMs like ChatGPT predominantly function to support medical knowledge, and while this knowledge is a lynchpin for the broader competencies of the physician, it is not the entirety of clinical practice or the learning expected of the medical student transforming into a student doctor and finally physician. In the clinical phase of medical education, where communication and procedural skills rise to prominence, the medical knowledge supported by LLMs cannot meet the patient-centered values and ethical considerations required for human interaction in the hospital. As with existing medical knowledge bases and clinical decision support (eg, UpToDate or DynaMedex), LLMs can be valuable adjuncts to clinical education. It is critical that LLMs do not detract from the humanistic elements of practice that are developed through clinical education.

Future Integration of LLMs Into Health Care and the Importance of Understanding Strengths and Weaknesses

The integration of LLMs into health care is fast becoming a reality, with both the availability of LLMs at students' fingertips and the rapid influx of research-driven deployments. Such integration is underscored by the impending inclusion of

ChatGPT into Epic Systems Corporation's software [36]. Potential applications range from reducing administrative tasks, like generating patient discharge instructions, assisting with insurance filings, and obtaining prior authorizations for medical services [37], to improving care quality through extracting key past medical history from complex patient records and providing interactive cross-checks of standard operating procedures (Figure 6).

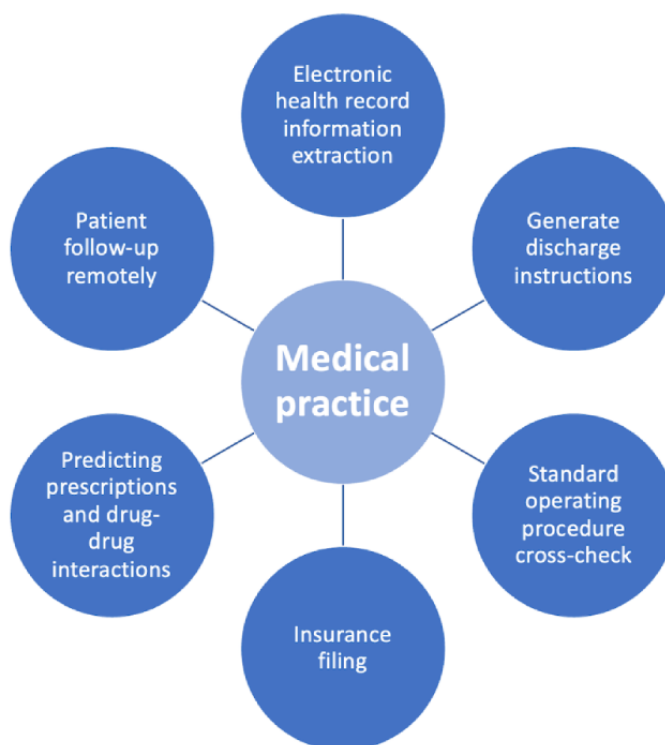
Across the range of emerging applications, the most notable are the potential for LLMs to digest the huge volumes of unstructured data in electronic health records and the possibility for LLMs to assist with clinical documentation [9,38]. However, these benefits are not without their challenges. Ethical considerations must be addressed regarding the impacts of misinformation and bias if LLMs are implemented to help generate clinical notes or instructions for patients or if they are applied to automate chart review for clinical research. Systematic approaches and ethical frameworks must be developed to mitigate these risks. Moreover, steps must be taken

to ensure that the use of patients' protected health information is in accordance with the Health Insurance Portability and Accountability Act (HIPAA) privacy and security requirements.

As we move toward a health care landscape increasingly intertwined with artificial intelligence, medical students must become adept at understanding and navigating the strengths and weaknesses of such technologies [39-41]. To be future leaders in health care, we must critically evaluate the best ways to harness artificial intelligence for improving health care while being cognizant of its limitations and the ethical, legal, and practical challenges it may pose.

The proactive curricular discourse surrounding topics like hallucinations, bias, and artificial intelligence models' self-evaluation of uncertainty, coupled with an exploration of potential legal and ethical issues, might be woven into the delivery of topics related to physicians' responsibility. By readily encouraging these dialogues, students can prepare for the challenges and opportunities that will come with the future integration of artificial intelligence into health care.

Figure 6. A few examples of how ChatGPT may be integrated into health care, derived from current news sources and research projects within the clinical informatics community.



Conclusions

LLMs like ChatGPT hold significant potential for augmenting medical education. By integrating them into the educational process, we can foster critical thinking, promote creativity, and offer novel learning opportunities. Moreover, a deeper understanding of these models prepares students for their

impending role in a health care landscape increasingly intertwined with artificial intelligence. Reflecting on the use of ChatGPT in medical school is an essential step to harness the potential of technology to lead the upcoming transformations in the digital era of medicine. The next generation of health care professionals must be not only conversant with these technologies but also equipped to leverage them responsibly and effectively in the service of patient care.

Acknowledgments

Research reported in this publication was supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health under award T35HL007649 (CWS), the National Institute of General Medical Sciences of the National Institutes of

Health under award T32GM136651 (AESE), the National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health under award T35DK104689 (AG), and the Yale School of Medicine Fellowship for Medical Student Research (AG). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Authors' Contributions

CWS, AESE, and DC contributed to the study conceptualization and drafting of the original manuscript. All authors participated in the investigation and validation process. All authors edited the manuscript draft and reviewed the final manuscript.

Conflicts of Interest

None declared.

References

1. Introducing ChatGPT. OpenAI. URL: <https://openai.com/blog/chatgpt> [accessed 2023-06-06]
2. Brants T, Popat AC, Xu P, Och FJ, Dean J. Large language models in machine translation. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 2007 Presented at: EMNLP-CoNLL; June 2007; Prague p. 858-867
3. Singh S, Mahmood A. The NLP cookbook: modern recipes for transformer based deep learning architectures. *IEEE Access* 2021;9:68675-68702 [doi: [10.1109/access.2021.3077350](https://doi.org/10.1109/access.2021.3077350)]
4. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, et al. Training language models to follow instructions with human feedback. In: Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A, editors. *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*. La Jolla, CA: Neural Information Processing Systems Foundation, Inc; 2022:27730-27744
5. Hirschberg J, Manning CD. *Advances in natural language processing*. *Science* 2015 Jul 17;349(6245):261-266 [doi: [10.1126/science.aaa8685](https://doi.org/10.1126/science.aaa8685)] [Medline: [26185244](https://pubmed.ncbi.nlm.nih.gov/26185244/)]
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 06;9:e46885 [FREE Full text] [doi: [10.2196/46885](https://doi.org/10.2196/46885)] [Medline: [36863937](https://pubmed.ncbi.nlm.nih.gov/36863937/)]
7. Lee H. The rise of ChatGPT: exploring its potential in medical education. *Anat Sci Educ* 2023 Mar 14:1 [doi: [10.1002/ase.2270](https://doi.org/10.1002/ase.2270)] [Medline: [36916887](https://pubmed.ncbi.nlm.nih.gov/36916887/)]
8. Xue VW, Lei P, Cho WC. The potential impact of ChatGPT in clinical and translational medicine. *Clin Transl Med* 2023 Mar;13(3):e1216 [FREE Full text] [doi: [10.1002/ctm2.1216](https://doi.org/10.1002/ctm2.1216)] [Medline: [36856370](https://pubmed.ncbi.nlm.nih.gov/36856370/)]
9. 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):1233-1239 [doi: [10.1056/NEJMSr2214184](https://doi.org/10.1056/NEJMSr2214184)] [Medline: [36988602](https://pubmed.ncbi.nlm.nih.gov/36988602/)]
10. OpenAI. GPT-3 model card. GitHub. 2022 Sep 01. URL: <https://github.com/openai/gpt-3/blob/master/model-card.md#data> [accessed 2023-06-23]
11. Olson P. ChatGPT needs to go to college. Will OpenAI pay? *The Washington Post*. 2023 Jun 05. URL: https://www.washingtonpost.com/business/2023/06/05/chatgpt-needs-better-training-data-will-openai-and-google-pay-up-for-it/f316828c-035d-11ee-b74a-5bdd335d4fa2_story.html [accessed 2023-06-20]
12. Barr K. GPT-4 is a giant black box and its training data remains a mystery. *Gizmodo*. 2023 Mar 16. URL: <https://gizmodo.com/chatbot-gpt4-open-ai-ai-bing-microsoft-1850229989> [accessed 2023-06-23]
13. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare (Basel)* 2023 Mar 19;11(6):887 [FREE Full text] [doi: [10.3390/healthcare11060887](https://doi.org/10.3390/healthcare11060887)] [Medline: [36981544](https://pubmed.ncbi.nlm.nih.gov/36981544/)]
14. Farrell L, Bourgeois-Law G, Regehr G, Ajjawi R. Autoethnography: introducing 'I' into medical education research. *Med Educ* 2015 Oct;49(10):974-982 [doi: [10.1111/medu.12761](https://doi.org/10.1111/medu.12761)] [Medline: [26383069](https://pubmed.ncbi.nlm.nih.gov/26383069/)]
15. Basic medical education WFME global standards for quality improvement: the 2020 revision. *World Federation for Medical Education*. 2020. URL: <https://wfme.org/wp-content/uploads/2020/12/WFME-BME-Standards-2020.pdf> [accessed 2023-06-23]
16. Wijnen-Meijer M, Burdick W, Alofs L, Burgers C, ten Cate O. Stages and transitions in medical education around the world: clarifying structures and terminology. *Med Teach* 2013 Apr;35(4):301-307 [doi: [10.3109/0142159X.2012.746449](https://doi.org/10.3109/0142159X.2012.746449)] [Medline: [23360484](https://pubmed.ncbi.nlm.nih.gov/23360484/)]
17. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. *arXiv Preprint posted online on December 26, 2022*. [FREE Full text] [doi: [10.1038/s41586-023-06455-0](https://doi.org/10.1038/s41586-023-06455-0)]
18. Xu G, Rong W, Wang Y, Ouyang Y, Xiong Z. External features enriched model for biomedical question answering. *BMC Bioinformatics* 2021 May 26;22(1):272 [FREE Full text] [doi: [10.1186/s12859-021-04176-7](https://doi.org/10.1186/s12859-021-04176-7)] [Medline: [34039273](https://pubmed.ncbi.nlm.nih.gov/34039273/)]
19. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. *J Am Med Inform Assoc* 2011;18(5):544-551 [FREE Full text] [doi: [10.1136/amiajnl-2011-000464](https://doi.org/10.1136/amiajnl-2011-000464)] [Medline: [21846786](https://pubmed.ncbi.nlm.nih.gov/21846786/)]

20. Johnson D, Goodman R, Patrinely J, Stone C, Zimmerman E, Donald R, et al. Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT model. Res Square. Preprint posted online on February 28, 2023. [FREE Full text] [doi: [10.21203/rs.3.rs-2566942/v1](https://doi.org/10.21203/rs.3.rs-2566942/v1)] [Medline: [36909565](https://pubmed.ncbi.nlm.nih.gov/36909565/)]
21. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, 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 Feb 08;9:e45312 [FREE Full text] [doi: [10.2196/45312](https://doi.org/10.2196/45312)] [Medline: [36753318](https://pubmed.ncbi.nlm.nih.gov/36753318/)]
22. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198 [FREE Full text] [doi: [10.1371/journal.pdig.0000198](https://doi.org/10.1371/journal.pdig.0000198)] [Medline: [36812645](https://pubmed.ncbi.nlm.nih.gov/36812645/)]
23. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. arXiv. Preprint posted online on March 20, 2023. [FREE Full text]
24. The core entrustable professional activities (EPAs) for entering residency. Association of American Medical Colleges. URL: <https://www.aamc.org/about-us/mission-areas/medical-education/cbme/core-epas> [accessed 2023-06-23]
25. Core entrustable professional activities. School of Medicine, Vanderbilt University. 2019. URL: <https://medschool.vanderbilt.edu/md-gateway/core-entrustable-professional-activities/> [accessed 2023-06-23]
26. GPT best practices. OpenAI. URL: <https://platform.openai.com/docs/guides/gpt-best-practices> [accessed 2023-06-27]
27. Hrynchak P, Batty H. The educational theory basis of team-based learning. Med Teach 2012;34(10):796-801 [doi: [10.3109/0142159X.2012.687120](https://doi.org/10.3109/0142159X.2012.687120)] [Medline: [22646301](https://pubmed.ncbi.nlm.nih.gov/22646301/)]
28. Charlin B, Roy L, Brailovsky C, Goulet F, van der Vleuten C. The Script Concordance test: a tool to assess the reflective clinician. Teach Learn Med 2000;12(4):189-195 [doi: [10.1207/S15328015TLM1204_5](https://doi.org/10.1207/S15328015TLM1204_5)] [Medline: [11273368](https://pubmed.ncbi.nlm.nih.gov/11273368/)]
29. Croskerry P. Achieving quality in clinical decision making: cognitive strategies and detection of bias. Acad Emerg Med 2002 Nov;9(11):1184-1204 [FREE Full text] [doi: [10.1111/j.1553-2712.2002.tb01574.x](https://doi.org/10.1111/j.1553-2712.2002.tb01574.x)] [Medline: [12414468](https://pubmed.ncbi.nlm.nih.gov/12414468/)]
30. Jones E, Steinhardt J. Capturing failures of large language models via human cognitive biases. arXiv. Preprint posted online on February 24, 2022. [FREE Full text]
31. Kost A, Chen FM. Socrates was not a pimp: changing the paradigm of questioning in medical education. Acad Med 2015 Jan;90(1):20-24 [doi: [10.1097/ACM.0000000000000446](https://doi.org/10.1097/ACM.0000000000000446)] [Medline: [25099239](https://pubmed.ncbi.nlm.nih.gov/25099239/)]
32. Hersh W, Ehrenfeld J. Clinical informatics. In: Skochelak SE, editor. Health Systems Science. 2nd Edition. Amsterdam, The Netherlands: Elsevier; May 06, 2020:105-116
33. Asher R. Clinical sense. The use of the five senses. Br Med J 1960 Apr 02;1(5178):985-993 [FREE Full text] [doi: [10.1136/bmj.1.5178.985](https://doi.org/10.1136/bmj.1.5178.985)] [Medline: [13794723](https://pubmed.ncbi.nlm.nih.gov/13794723/)]
34. Talley N, O'Connor S. Clinical Examination: A Systematic Guide to Physical Diagnosis. Amsterdam, The Netherlands: Elsevier; 2014.
35. Martin A, Weller I, Amsalem D, Duvivier R, Jaarsma D, de Carvalho Filho MA. Co-constructive patient simulation: a learner-centered method to enhance communication and reflection skills. Simul Healthc 2021 Dec 01;16(6):e129-e135 [FREE Full text] [doi: [10.1097/SIH.0000000000000528](https://doi.org/10.1097/SIH.0000000000000528)] [Medline: [33273424](https://pubmed.ncbi.nlm.nih.gov/33273424/)]
36. Adams K. Epic to integrate GPT-4 into its EHR through expanded Microsoft partnership. MedCity News. 2023. URL: <https://medcitynews.com/2023/04/epic-to-integrate-gpt-4-into-its-ehr-through-expanded-microsoft-partnership/> [accessed 2023-06-20]
37. Landi H. Doximity rolls out beta version of ChatGPT tool for docs aiming to streamline administrative paperwork. Fierce Healthcare. 2023. URL: <https://www.fiercehealthcare.com/health-tech/doximity-rolls-out-beta-version-chatgpt-tool-docs-aiming-streamline-administrative> [accessed 2023-06-21]
38. Landi H. Microsoft's Nuance integrates OpenAI's GPT-4 into voice-enabled medical scribe software. Fierce Healthcare. 2023. URL: <https://www.fiercehealthcare.com/health-tech/microsofts-nuance-integrates-openais-gpt-4-medical-scribe-software> [accessed 2023-06-27]
39. Chartash D, Rosenman M, Wang K, Chen E. Informatics in undergraduate medical education: analysis of competency frameworks and practices across North America. JMIR Med Educ 2022 Sep 13;8(3):e39794 [FREE Full text] [doi: [10.2196/39794](https://doi.org/10.2196/39794)] [Medline: [36099007](https://pubmed.ncbi.nlm.nih.gov/36099007/)]
40. Hersh WR, Gorman PN, Biagioli FE, Mohan V, Gold JA, Mejicano GC. Beyond information retrieval and electronic health record use: competencies in clinical informatics for medical education. Adv Med Educ Pract 2014;5:205-212 [FREE Full text] [doi: [10.2147/AMEP.S63903](https://doi.org/10.2147/AMEP.S63903)] [Medline: [25057246](https://pubmed.ncbi.nlm.nih.gov/25057246/)]
41. Paranjape K, Schinkel M, Nannan Panday R, Car J, Nanayakkara P. Introducing artificial intelligence training in medical education. JMIR Med Educ 2019 Dec 03;5(2):e16048 [FREE Full text] [doi: [10.2196/16048](https://doi.org/10.2196/16048)] [Medline: [31793895](https://pubmed.ncbi.nlm.nih.gov/31793895/)]

Abbreviations

- EPA:** Entrustable Professional Activity
- HIPAA:** Health Insurance Portability and Accountability Act
- LLM:** large language model
- USMLE:** United States Medical Licensing Exam

Edited by T de Azevedo Cardoso; this is a non-peer-reviewed article. Submitted 17.07.23; accepted 26.07.23; published 14.08.23.

Please cite as:

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;9:e50945

URL: <https://mededu.jmir.org/2023/1/e50945>

doi: [10.2196/50945](https://doi.org/10.2196/50945)

PMID: [37578830](https://pubmed.ncbi.nlm.nih.gov/37578830/)

©Conrad W Safranek, Anne Elizabeth Sidamon-Eristoff, Aidan Gilson, David Chartash. Originally published in JMIR Medical Education (<https://mededu.jmir.org>), 14.08.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on <https://mededu.jmir.org/>, as well as this copyright and license information must be included.