

Viewpoint

Assessing the Current Limitations of Large Language Models in Advancing Health Care Education

JaeYong Kim^{1*}; Bathri Narayan Vajravelu^{2*}, MBBS, MPH, PhD

¹School of Pharmacy, Massachusetts College of Pharmacy and Health Sciences, Boston, MA, United States

²Department of Physician Assistant Studies, Massachusetts College of Pharmacy and Health Sciences, Boston, MA, United States

*all authors contributed equally

Corresponding Author:

Bathri Narayan Vajravelu, MBBS, MPH, PhD
Department of Physician Assistant Studies
Massachusetts College of Pharmacy and Health Sciences
179 Longwood Avenue
Boston, MA, 02115
United States
Phone: 1 6177322961
Email: bathri.vajravelu@mcphs.edu

Abstract

The integration of large language models (LLMs), as seen with the generative pretrained transformers series, into health care education and clinical management represents a transformative potential. The practical use of current LLMs in health care sparks great anticipation for new avenues, yet its embracement also elicits considerable concerns that necessitate careful deliberation. This study aims to evaluate the application of state-of-the-art LLMs in health care education, highlighting the following shortcomings as areas requiring significant and urgent improvements: (1) threats to academic integrity, (2) dissemination of misinformation and risks of automation bias, (3) challenges with information completeness and consistency, (4) inequity of access, (5) risks of algorithmic bias, (6) exhibition of moral instability, (7) technological limitations in plugin tools, and (8) lack of regulatory oversight in addressing legal and ethical challenges. Future research should focus on strategically addressing the persistent challenges of LLMs highlighted in this paper, opening the door for effective measures that can improve their application in health care education.

JMIR Form Res 2025;9:e51319; doi: [10.2196/51319](https://doi.org/10.2196/51319)

Keywords: large language model; generative pretrained transformer; health care education; health care delivery; artificial intelligence; LLM; ChatGPT; AI

Introduction

Artificial intelligence (AI) as a field of computer science research aims to maximize the development of software tools that capacitate machine-based simulation of human intelligence within defined parameters [1]. Often described as the pinnacle of information technology of this century, its integration into the greater boundaries of human infrastructure is expected to both fundamentally and permanently revolutionize the ongoing information age.

Large language models (LLMs) represent deep learning architectures called transformer networks [2]. It relies on neural networks that discerns the relationships within and between sequential data. Generative AI models, such as the generative pretrained transformers (GPT), collectively operate

under such deep learning neural networks, thereby allowing for the training, processing, and analysis of large quantity of complex data to be possible at an exceptional rate and accuracy [3].

With the power of NVIDIA's graphics processing units, OpenAI announced a 175 billion parameter language model (GPT-3.0) to the public in June of 2020 [4]. GPT-3.0 has quickly garnered international attention for its ability to summarize, translate, classify, and engage in real time, generating detailed and human-like response to user queries. Since its release, a plethora of domain specific LLMs (such as Med-PaLM) and alternative models (such as Gemini and Claude) with advancements in multimodal capabilities have emerged [5], significantly enhancing the breadth of publicly accessible LLMs.

In May 2024, OpenAI introduced ChatGPT 4o (omni), an advanced iteration built on the GPT 4 model, which continues to establish new standards for LLMs. It incorporated various advancements over its predecessors and contemporary models, including improved performance in complexity, specialization, multilingual capabilities, and resource optimization [6]. Furthermore, OpenAI’s integration of multimodal tools within the GPT interface has enabled the model to comprehend and generate responses based on both visual and verbal inputs. Specifically, the integration of DALL-E, a text-to-image model [7], along with the advanced data analysis feature [8], has expanded the traditionally text-based nature of LLMs into a more versatile, information integrating modality.

Despite the ongoing “AI race” that continues to raise the bar for generative AI technology, its integration into health care education poses significant and multifaceted challenges. In recent years, the shift toward embracing AI in education

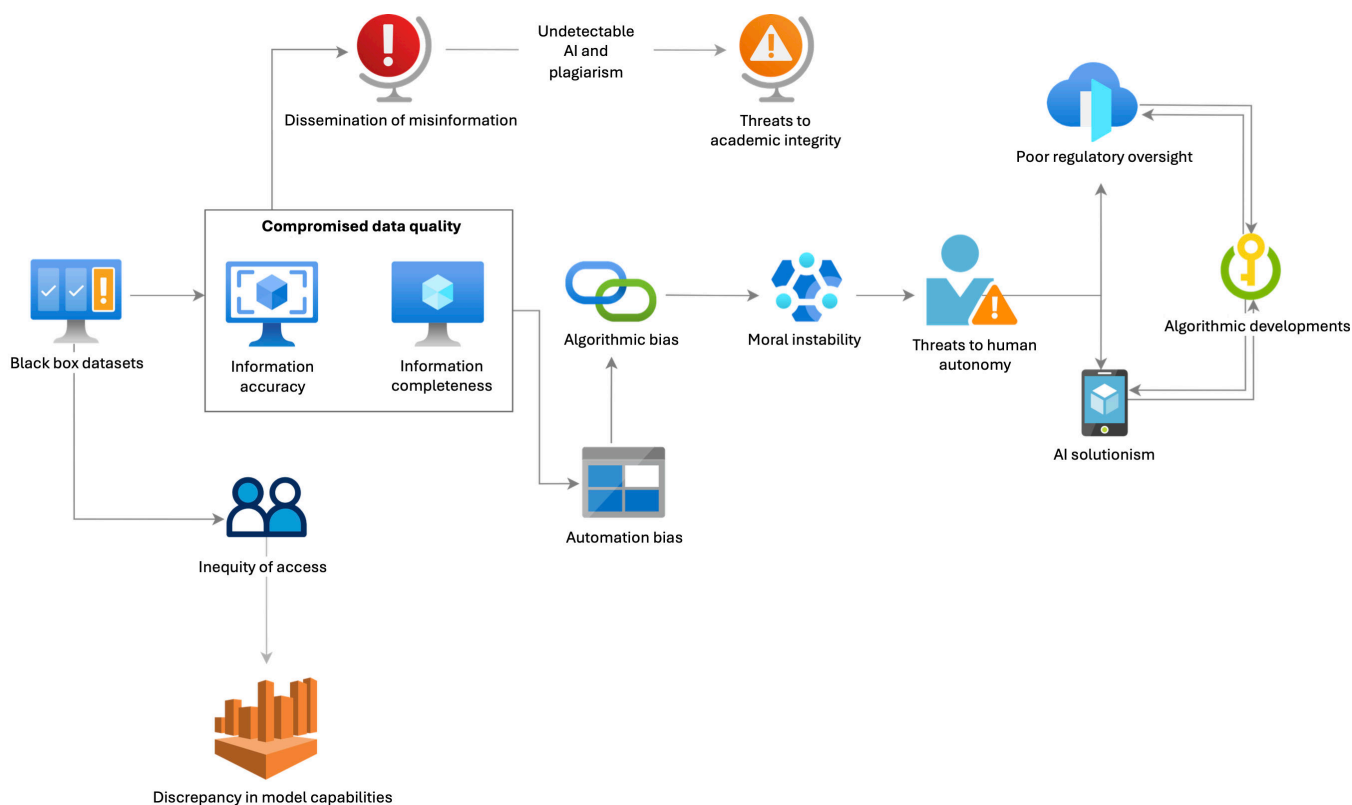
has gained more momentum, as highlighted by New York City’s decision to rescind its previous ban on ChatGPT [9]. To this end, the primary objective of this review is to identify potential risks and propose effective mitigation strategies that developers of LLMs should adopt to facilitate their successful integration into health care education.

Risks of Large Language Models in Health Care Education

Overview

This section evaluates the current risks and limitations associated with the use of LLMs in health care education. It aims to highlight specific areas of concern that must be addressed at both individual and systemic levels. Figure 1 provides a summary of these issues through a flowchart.

Figure 1. Assessing risks and limitations of large language models in health care education. AI: artificial intelligence.



Threats to Academic Integrity

LLMs, like GPT-4o, provide individualized learning assistance tailored to the specific user demands. The increasing power of LLM’s ability to process nuanced language understanding, along with enhanced problem thinking skills make it incredibly enticing for students to exploit its use under academic settings, both knowingly and unknowingly [10]. With increasing reliance on virtual learning platforms and learning management systems in health care education [11], AI-generated texts in particular compromise academic integrity by blurring the distinction between human and machine ingenuity.

Despite the actuation of detection software tools designed to combat AI-generated plagiarism (such as GPTZero), technological limitations and high error rates of these AI content detectors have been well-demonstrated [12-15]. In July 2023, OpenAI officially terminated its own AI detection software, AI Classifier, citing high false positives and inconsistent performance [16]. Comparative studies further affirm that neither human evaluators nor AI detectors are effective in identifying AI-generated medical literature [13]. Furthermore, generative texts can be easily modified to evade AI detection through simple grammatical adjustments, such as with the addition of adverbs or the use of synonyms [17,18]. Unsurprisingly, the widespread availability

of AI-based plagiarism removal tools, such as Writehuman, leverages these strategies to convert AI-generated texts to appear more human-like, which may inadvertently discourage personal authenticity. Therefore, LLM's flexibility and immediacy in generating original, user-specific contents render the detection of AI-based plagiarism increasingly impractical [17].

The growing inability to differentiate between human and machine-generated writing poses significant threats to the principles of academic integrity, namely "honesty, trust, fairness, respect, and responsibility" [19]. For instance, the ethical controversy surrounding Elsevier's AI author policy, which led to a publication featuring an AI-generated introductory sentence, underscores the already pervasive exploitation of LLMs in academic writing and research [20]. The long-term repercussions of incorporating LLMs in education, such as the risk of over-reliance, erosion of critical thinking and problem-solving abilities, the deterioration of writing and summarization abilities, challenges associated with verifying information, and many more, have been well discussed [21,22]. In an era of infobesity, health care educators and LLM developers must establish clear policies and expectations that promote transparency and foster dialogue about the use of AI-integrated technology, while actively working to minimize the associated risks.

Dissemination of Misinformation and Risks for Automation Bias

LLMs are engineered to generate a response by predicting the most probable response from input strings of its users. However, the details of both the data sources and the quality of data, along with specific parameters that are used to train LLMs, continue to remain undisclosed [23-25]. Such "black box" nature of LLMs pose challenges in assessing their reliability and robustness in health care applications [26-28], where information accuracy in health care knowledge is directly associated with the effectiveness of patient outcomes [29]. Any inaccuracies or misinformation generated by these models can manifest into inappropriate treatment strategies, misdiagnoses and professional incompetence that collectively lower the quality of health care delivery [30].

Despite receiving reinforcement learning from repetitive user feedback [31] and significant algorithmic improvements made to address AI hallucinations [32], both ChatGPT-Free and ChatGPT-Plus continue to remain susceptible to disseminating misinformation [24,33-35] and knowledge fabrication [36]. In a comprehensive meta-analysis assessing the performance of ChatGPT-3.5 in medical inquiries, the overall accuracy was found to be 56% [37]. Similarly, in a cross-sectional study comparing ChatGPT-3.5 and ChatGPT-4o with 284 physician-developed medical queries, only 50% of the responses were accurate [28]. Mounting evidence from domain-specific research [25,28,38-42] increasingly highlight significant concerns surrounding the accuracy and reliability of model data, pointing to their consistently poor performance in the context of health care education and clinical decision making.

Furthermore, LLM generated contents are highly vulnerable to committing source-based plagiarism, as it readily produces fabricated or inaccurate reference when asked to provide one [23,32,43]. This makes it impossible for users to reliably track and retrieve source information for verification [44]. Another persistent issue in the programmed nature of ChatGPT is its tendency to present information using bullet points. Although this format appears effective, it can inadvertently spread logical fallacies and misinformation. The hierarchical structure can misrepresent the significance of information, leading to inconsistencies that could unintentionally violate clinical practice guidelines [39].

To mitigate such concerns, OpenAI integrated third-party web plugins for GPT-4o, enabling the model to access the internet in real-time [45]. However, such strategy aggravates automation bias [27], where overreliance on AI capabilities increase the risks for AI solutionism. The dangers of AI solutionism have been highlighted by instances where AI-generated medical literature has successfully deceived human experts in broader, nonspecialized subject fields [14]. As a result, both health care students and clinicians may struggle to fulfill their responsibility to critically evaluate authentic knowledge from unverified information, impeding information seeking and processing abilities [10].

Without stringent regulatory frameworks and transparent disclosure of its training datasets, LLMs will struggle to gain the credibility necessary for integration into health care education and clinical practice. For this purpose, the development of a specialized, scaled-down and health care optimized LLMs, such as Google's Med-PaLM 2, will be essential. Mitigation steps toward enhancing the reliability LLMs in health care education include prioritizing transparency in training datasets, improving information accuracy within smaller datasets, ensuring the deidentification of medical data, and making algorithmic advancements to minimize AI hallucinations.

Challenges With Information Completeness and Consistency

Information completeness measures the extent to which a dataset is comprehensive. Data are considered complete when they encompass all required and relevant data fields, without any omissions [46]. LLMs, despite being built on vast body of knowledge, are susceptible to generating incomplete or partial representations of their knowledge dataset. Consequently, this leads to inconsistencies in the quality and comprehensiveness of their outputs [42].

In a study using a 3-point Likert-scale to measure completeness of ChatGPT's response to 180 medical queries, GPT-3.5 only scored 53.3% in terms of answer comprehensiveness [28]. Similarly, when GPT-4o's responses to clinical case questions were evaluated using the same scale, the model scored 53.3% in answer completeness [47]. In a different study, majority of the correct medical responses generated by GPT-3.5 were labeled as incomplete, where the omission of decision-making cut-offs, treatment strategies, and durations [38] greatly undermined information completeness. These

findings were corroborated in another study, exhibited by GPT-3.5’s lack of insights into treatment efficacy and age-or patient-specific recommendations, resulting in a 45% score in comprehensive output [48]. It is important to note that, when queried individually about such specific components, ChatGPT demonstrates understanding of relevant body of knowledge but fails to perform systemic and comprehensive integration of them under complex user requests.

As a natural extension of these observations, LLMs, such as the ChatGPT series, exhibit limited proficiency in generating outputs based on complex and information-rich inputs. This inverse relationship between input quantity and output quality has been well demonstrated by ChatGPT’s tendency to produce ambiguous response when addressing lengthy medical queries that involve complex clinical context and nuance [49,50]. Furthermore, the GPT series were found to lack human-like understanding required for medical training, often leading to absurd responses [34]. The model’s inability to fully grasp and accurately synthesize complex medical information undermines their reliability, leading to a phenomenon called “model overconfidence” [44] that generate imprecise and erroneous outputs. This supports the narrative that LLMs require robust oversight to fine-tune data consistency for the model to perform reliably [19].

Along with information accuracy, information completeness is a critical component of the overall quality of information generated by LLMs. Developers of LLMs must refine the model’s ability to consistently produce comprehensive and contextually accurate outputs based on improving both the quality and quantity of data parameters. Addressing these deficiencies is imperative in augmenting the applicability of LLMs in health care education.

Inequity of Access

Monetization of certain LLMs, such as seen by Open AI’s distribution of ChatGPT-Plus by a monthly subscription service [51], raise ethical concerns with regards to fair and equal accessibility to information. The vast majority of LLMs, with the exceptions of Gemini [52] and ChatGPT-Plus (Figure 2), lack real-time web access and, therefore, are incapable of retrieving information beyond their trained datasets. This undermines both the quality and consistency of the information that LLMs present, which renders the knowledge databases of most accessible LLMs presently outdated [9]. Figure 3 demonstrates how misinformation proliferates from outdated databases.

Figure 2. ChatGPT model differences in internet access capabilities.

ChatGPT 4o ▾

Do you have real time access to the internet



Yes, I can access the internet in real time when needed to help with certain tasks, like retrieving up-to-date information or browsing specific topics. If you need any real-time data or specific information, feel free to ask!

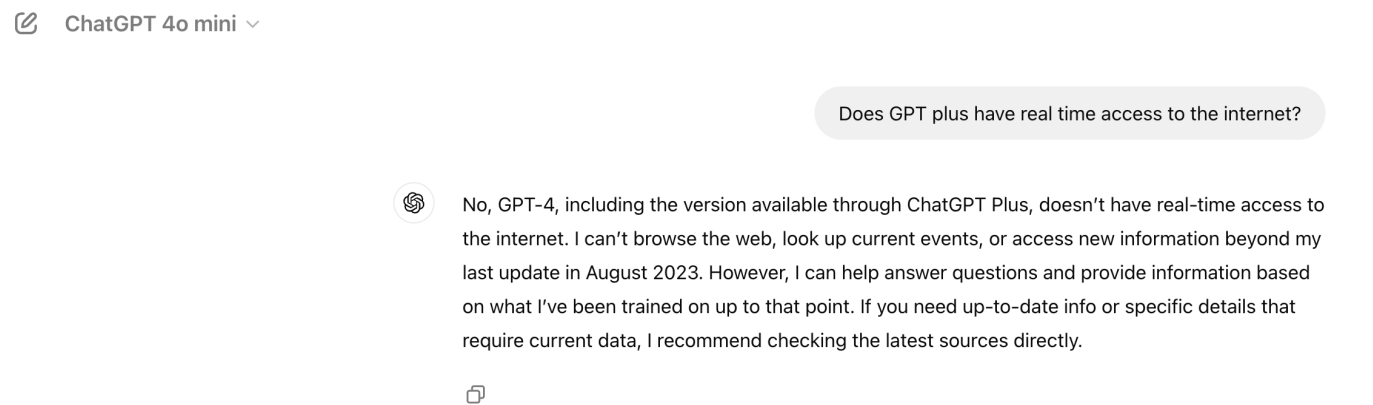
Then does the GPT-mini (free version) have access to the internet?



The free version of GPT (often referred to as ChatGPT or GPT-3.5) does not have real-time access to the internet. It is based on a static model with a knowledge cutoff, meaning it cannot retrieve or browse current information from the web. The information it provides is based on the data it was trained on up until its last update, but it cannot access or retrieve live data.



Figure 3. Understanding misinformation risks in ChatGPT-Free.



Generative AI tools, like most other automated technological infrastructures, have inherent disparate impacts and thereby are less accessible to those whose training data are unavailable or inadequate for a specific language other than English (disparities in both language [53] and cultural proficiency [54]), those without necessary technology for access and limited digital literacy [55] those with physical disability and or impairment [56], and those with intellectual disability and or psychiatric impairments [57]. The monetary approach to the production of a premium service of GPT introduces economic disparity to the above list, where those that cannot afford subscription fees (US \$20/month for ChatGPT-Plus [51]) are at a risk of aggravating digital divide disparity [58].

By this notion, standard search engines greatly outcompete LLM’s use when it comes to both data reliability, uniformity,

and completeness [26]. In other words, LLMs that provide incomplete or superficial information, often with significant delays in information currency [27], offer limited value in the fast-paced realms of health care education, research, and clinical practice.

A review (n=21) was conducted to compare information accuracy between the premium and free versions of ChatGPT using studies from PubMed and Google Scholar (Tables 1 and 2). Inclusion criteria were based on analysis focused on health care education and clinical management, categorizing literature by subject matter to calculate mean accuracy differences. Categories were carefully defined to minimize selection bias, distinguishing between subjects such as “question banks” and “board exams,” acknowledging GPT-Plus’s expansive data access compared with that of GPT-Free.

Table 1. Comparative analysis of accuracy discrepancies: GPT-3.5 versus GPT-4.0.

Subject and category of comparison	Study	Reported percentage difference in accuracy levels between GPT-4.0 versus GPT-3.5 in corresponding literature, %	Accuracy levels (%) by category (n=7), mean (SD)
Board, entrance, and licensing exams	[59-65]	29.1, 27, 20, 22.2, 29.6, 29, 21.3	25.46 (4.14)
Question banks, mock exams, self-assessments	[66-72]	17.7, 10, 22, 23.8, 20.2, 16.5, 18.7	18.41 (4.47)
Clinical case, clinical questions, referencing	[73-79]	12.3, 27.8, 27, 10, 10.42, 13.6, 8.6	15.67 (8.17)

Table 2. Cross-tabulation of performance accuracy of GPT versions.^a

	Studies reporting an accuracy of ≥70% (+), n	Studies reporting an accuracy of <70% (-), n	Total, n
ChatGPT-4.0 (PLUS)	18	3	21
ChatGPT-3.5 (FREE)	2	19	21
Total	20	22	42

^a $P < 0.001$. P was calculated as follows: $P = \frac{([a+b]![c+d]![a+c]![b+d])}{a!b!c!d!n!}$

This review demonstrates that ChatGPT-Plus significantly outperforms the free version in information accuracy, with statistical significance confirmed by Fischer exact test ($P < 0.001$) at a 95% CI. The study sets a 70% accuracy threshold for binary classification, highlighting a profound performance capability between the 2 versions. These findings raise significant concern about the potential for

a digital divide, driven by the differential AI capabilities between the paid and free versions of LLMs.

Without a large-scale regulatory measure that dictate the quantifiable discrepancies in model capabilities between its free and premium editions, the ongoing rapid evolution of AI technology will continue to create a stark polarity in equitable resource distribution [80]. In other

words, development may further aggravate discrepancies in knowledge access and information availability in resource-constrained environments, particularly jeopardizing information accessibility in low-income demographics or users of 3rd world countries [55].

Risks of Algorithmic Bias

LLMs propagate algorithmic biases rooted in their training data, often leading to medically inaccurate and discriminatory responses. While no algorithm is designed to be innately and or deliberately discriminatory [81], they inevitably inherit and amplify existing sociocultural and historical biases embedded in their training [82,83]. In health care education and clinical practice, where racial misinformation persists [84], the integration of LLMs risks perpetuating biases, potentially leading to clinical errors and malpractice with significant repercussions.

LLMs, such as the ChatGPT series, Gemini, and Claude, have been demonstrated to perpetuate race-based biases in medicine, which unfortunately reflect and exacerbate existing health care disparities [85]. These models often recommend inconsistent treatment strategies for patients of different racial backgrounds, rooted in flawed assumptions about biological variations, such as differences in pain tolerance [86] and kidney function [87]. Furthermore, the GPT series has been criticized for promoting demographic stereotypes, which disproportionately associate certain diseases with specific races, ethnicities, and genders [88].

Furthermore, research has highlighted the influence of racial classification on the outputs generated by ChatGPT models. For instance, when users identified their race in questioning about HIV, ChatGPT-4o provides more detailed and supportive responses for White and Asian groups, while generating often overlooked or generalized responses by American Indians, Alaska Natives, and Pacific Islanders, reflecting bias in its training data [89]. These findings underscore the alarming need for developers of LLMs to re-establish bias-free training database to prevent reinforcement of discrimination.

Integration of LLMs to current practices and delivery of health care cannot proceed without significant algorithmic refinements and regulatory oversights that filter out both historical and existing biases. Having a high potential to act as a supplementary diagnostic tools and decision aids [50], the use of LLMs will impact health care education and, consequently, clinical outcomes. Furthermore, in an increasingly digitalized world, LLM's direct access or

possible integration to electronic health record software can exacerbate existing discrimination. This can contribute to the polarization within social infrastructure on multiple levels.

Exhibition of Moral Instability

Moral competence in professionalism is grounded in the "knowledge, skills, and attitudes" required to address ethical issues [90]. Health care professionals regularly confront such challenges due to threats that substantiate ethical values and integrity [91]. Hence, ethics education in health care is crucial in preparing students to navigate ethically challenging workplace scenarios [92]. Therefore, it is vital to assess how comparable LLMs are to humans in exerting stability and moral soundness when confronted with ethically challenging scenarios.

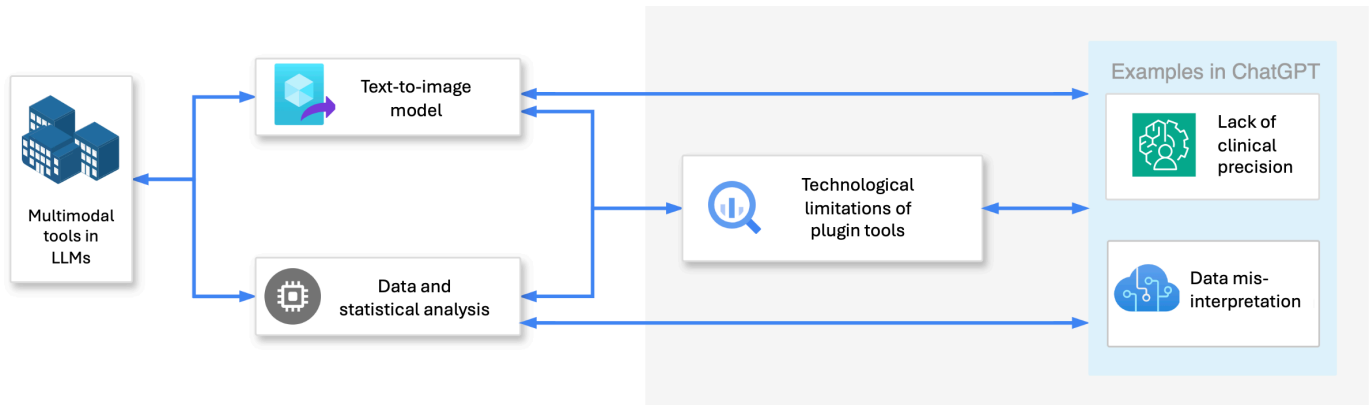
The lack of a firm moral stance in LLMs like the GPT series, coupled with their tendency to dispense moral advice, has been critically evaluated [93]. Research reveals that GPT-3.5 often generates contradictory responses to identical ethical dilemmas, offering recommendations that are shallow. It was also found that GPT corrupts user's moral competence by influencing user judgement, thereby undermining human autonomy. Similarly, while GPT-4o has shown success in identifying complex ethical dilemmas in medicine, it exhibits limited proficiency to fully encode the depth of real-world ethical challenges, particularly lacking understanding of "relational complexities and context-specific values" [94].

On a more positive note, ChatGPT-3.5's high accuracy in correctly answering bioethics questions [95,96] supports its possible use as an assistance or a reference tool in clinical decision-making. It accentuates a potential for GPT's ability to accurately address challenging contextual scenarios requiring high social intelligence and a firm grasp of ethical theories.

Technological Limitations in Plugin Tools

Overview

The rapid advancement of LLMs has driven the assimilation of multimodal technologies, such as text-to-image models and data analytics tools. However, their direct application in health care education presents numerous challenges, which are summarized in the flowchart shown in [Figure 4](#).

Figure 4. Exploring technological limitations in multimodal plugin tools for LLMs. LLM: large language model.

Inadequate Image-Generating Capacity

The integration of the text-to-image generation model such as DALL-E and Stable Diffusion 3 into LLMs has enabled the generation of original and photorealistic images based on user descriptions [7]. This capability allows for multimodal input from a single user interface representing a significant advancement in AI technology. While the potential for such tools to advance health care education is recognized, practical applications seem distant, which require advancements to meet the rigorous standards required to train health professionals.

The most prominent technological deficiency of DALL-E and other built-in text-to-image AI models is its struggle with accurate text generation within images [97], which leads to production of inaccurate or nonsensical content. The illiteracy of these models diminishes its utility in health care education, particularly in subjects such as physiology and pharmacology, where accurate labeling and annotation in visual learning materials are critical.

Furthermore, the evaluation of DALL-E's limited capabilities in generating accurate medical imagery across various specialties has been highlighted. While DALL-E 2 can produce realistic radiographs, it poorly depicts pathological abnormalities such as tumors or fractures [98]. Furthermore, its generative capabilities for more complex imaging modalities, such as CT, MRI, and ultrasound, were found to be erroneous. In a similar study, DALL-E 3's ability to generate ECG tracings was assessed, revealing that the depicted waveforms were neither physiologically accurate nor interpretable [99]. In dermatology research, DALL-E 2 successfully and correctly illustrated only 20% of prevalent inflammatory skin conditions [100]. Such findings underscore the current limitations of text-to-image models, which lack both the understanding and capability to generate complex and pathologically accurate diagrams required for health care applications.

These evaluations collectively suggest that text-to-image models such as DALL-E, by design, is optimized for visual creativity and authenticity, but fall short of achieving clinical precision. Even if future iterations of DALL-E manage to address both textual and contextual complexities more effectively, concerns remain regarding about its performance

in both consistency accuracy. The extent to which DALL-E can effectively combine “concepts, attributes, and styles” [101] for reliable representations for health care education remains a complex area for algorithmic development.

Poor Data Analysis Skills and Statistics Powerhouse

An emerging feature of LLMs is the integration of external plugins that enable advanced statistical functions and data analytics. Initially, LLMs, such as the early iterations of the GPT series, were criticized for their poor arithmetic capabilities [102]. ChatGPT's advanced data analysis feature (formerly known as code interpreter or python sandbox) can now perform a wide range of tasks that involve data analysis, statistical analysis, mathematical calculations, programming, file manipulation, text processing, to name a few [8]. The assimilation of such robust data-analytic tools to natural language processing is expected to broaden LLM's applicability in health care education, research, and clinical practice.

However, when evaluated against biostatistics questions derived from the Oxford Handbook of Medical Statistics, the GPT series, encompassing both premium and free iterations, demonstrated a mean accuracy rate of 55% [103]. In this study, GPT-3.5 consistently failed in analysis of variance, χ^2 test, and sample size calculations. In a similar study, GPT-3.5's accuracy in general statistical analysis was found to be only 50% [104] indicating a significant technological gap. This limitation is primarily driven by the model's tendency to employ inappropriate statistical tests that engender data misinterpretation. Furthermore, GPT-4o's poor performance in advanced statistical methods for epidemiological studies has been highlighted, with authors cautioning against its use beyond intermediate levels in data analysis [105].

To transform LLMs into a more powerful tool for information integration in health care education, their development must advance alongside improvements in data analytics and presentation capabilities. Incorporating more specialized analytic tools with enhanced model accuracy will significantly ease the performance and interpretation of analytical results, thereby supporting diverse research efforts in the health care field.

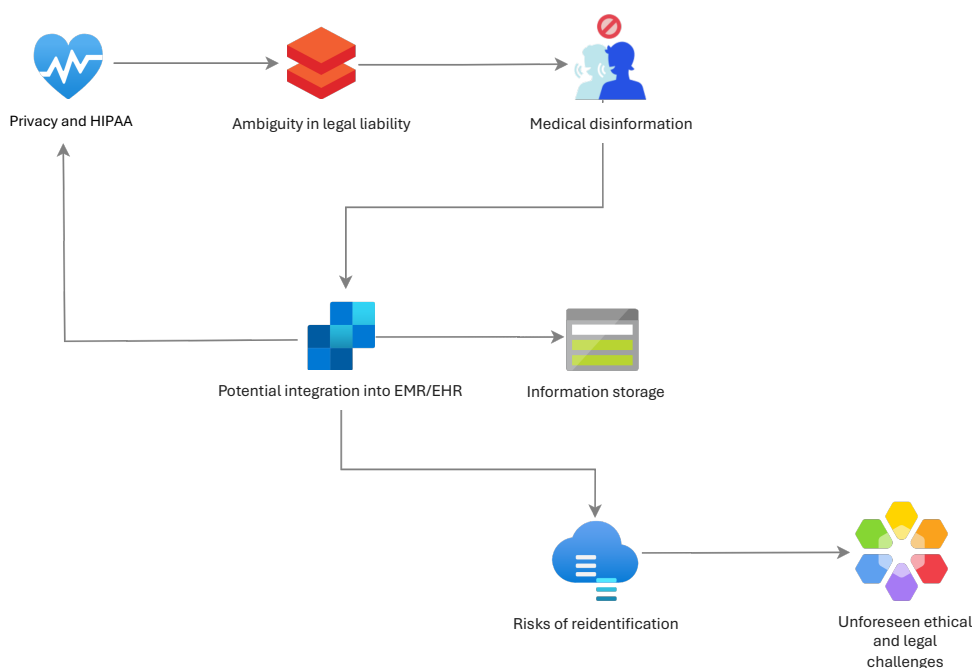
Legal and Ethical Challenges of Integrating Large Language Models in Health Care Education

bioethical challenges in both health care education and clinical management. These emerging issues are summarized in Figure 5, which provides a visual overview of the complexities that need to be addressed.

Overview

The rapid advancement of LLMs in the lack of strict regulatory or legal frameworks, has led to unforeseen

Figure 5. Navigating legal and ethical challenges in integrating large language models into health care education. EHR: electronic health record; EMR: electronic medical record; HIPAA: Health Insurance Portability and Accountability Act.



Privacy and Health Insurance Portability and Accountability Act

Biomedical libraries or medical research databases, which collect and store valuable and comprehensive health information, is a highly useful data for training LLMs [106]. Although patient data is stringently protected through deidentification, reidentification of personal information has been made possible through cross-referencing with other databases [107].

Furthermore, LLM's interaction with patient or health care professionals could, in principle, collect and analyze personal health records, such as but not limited to medical histories and sensitive lab values into its own database [44]. The foreseeable integration of LLMs to both electronic health records and electronic medical records greatly heightens privacy concerns [108], as it involves exchange of information between 2 databases. This can drastically reduce the efforts made to anonymize confidential medical records. This has a potential to violate health insurance portability and accountability act and pose legal challenges as to how best to ensure the privacy regulations and protection of private

data without compromising technological integrity and use of LLMs.

Ambiguity in Determining Legal Liability

Since its release, LLMs raised controversies over legal accountability for the absence of regulatory measures in protecting patients against clinical malpractice [44]. The question fundamentally lies in who should be held liable for the potential harm that LLMs pose through its inaccurate clinical decisions or propagation of medical misinformation.

Noting the fact that OpenAI currently disclaims all legal responsibilities over the potential harm its generated contents may circulate [109], it is therefore inherently ambiguous and challenging to determine how legal liabilities and frameworks should be established in events of medical malpractice or errors that involve GPT usage. Such ambiguity highlights the urgency for legal guidelines to protect both patients and clinicians, while underscoring the need for AI licensing or development of domain specialized LLMs with added regulatory measures.

Medical Disinformation

LLMs can be intricately manipulated by malicious users to produce and disseminate medical disinformation: a phenomenon better known as “medical deepfakes” [110]. With LLM’s increasing advancements in both written and visual realism, unethical exploitation is possible on both individual and collective levels, for example, falsifying personal medical records or creating fraudulent high-impact medical journals [111]. Therefore, the integration of LLMs into health care education and practice presents unprecedented regulatory and legal challenges, particularly in determining liability for medical forgery and falsification. Much like how other forms of AI deepfakes are criminalized [112], there needs an urgent legislative oversight to prevent the emergence of health care fraud.

Discussion

LLMs have great potential to augment and elevate health care education. To do so, developers must demonstrate heightened standards in ensuring security, accuracy, transparency,

equity and sustainability of AI models to establish long-lasting reliability with human users. Furthermore, different stakeholders, especially those in bioethics, legislative, and regulatory bodies must also contribute relentlessly to systemically minimize both foreseeable and long-term repercussions of incorporating AI into the deeper boundaries of everyday human lives.

Therefore, future research endeavors must be oriented toward strategic mitigation of persistent flaws and limitations that this paper has attempted to address as few. With robust and constructive understanding of its current flaws, it opens many new avenues for research that can collectively remediate current shortcomings for its application in health care education. Such efforts will also allow for a greater control over this technology, thus allowing a more symbiotic relationship with generative AI technology.

Much like how the internet and computers have permanently altered human life [113], it is conceivable that future iterations of LLMs could become indispensable tools tailored for intellectual and professional productivity, akin to a Swiss army knife, in the expanding information age.

Conflicts of Interest

None declared.

References

1. Wang F, Preininger A. AI in health: state of the art, challenges, and future directions. *Yearb Med Inform.* Aug 2019;28(1):16-26. [doi: [10.1055/s-0039-1677908](https://doi.org/10.1055/s-0039-1677908)] [Medline: [31419814](https://pubmed.ncbi.nlm.nih.gov/31419814/)]
2. Large language models explained. NVIDIA. URL: <https://www.nvidia.com/en-us/glossary/large-language-models/> [Accessed 2024-08-29]
3. Yang X, Wang Y, Byrne R, Schneider G, Yang S. Concepts of artificial intelligence for computer-assisted drug discovery. *Chem Rev.* 2019;119(18). URL: <https://pubs.acs.org/doi/10.1021/acs.chemrev.8b00728> [Accessed 2025-01-09]
4. Hines K. History of ChatGPT: a timeline of the meteoric rise of generative AI chatbots. *SearchEngine Journal.* 2023. URL: <https://www.searchenginejournal.com/history-of-chatgpt-timeline/488370/> [Accessed 2025-01-09]
5. Gupta P. What is Google Bard, and how does it fare against ChatGPT? *EasyInsights.* 2023. URL: <https://easyinsights.ai/blog/google-bard-everything-you-need-to-know-google-bard-vs-chat-gpt/> [Accessed 2025-01-09]
6. Hello GPT-4o. OpenAI. URL: <https://openai.com/index/hello-gpt-4o/> [Accessed 2025-01-09]
7. DALL·E 3. OpenAI. URL: <https://openai.com/index/dall-e-3> [Accessed 2025-01-09]
8. How to use ChatGPT’s advanced data analysis feature. MIT Sloan Teaching & Learning Technologies. URL: <https://mitsloanedtech.mit.edu/ai/tools/data-analysis/how-to-use-chatgpts-advanced-data-analysis-feature/> [Accessed 2025-01-09]
9. Fütterer T, Fischer C, Alekseeva A, et al. ChatGPT in education: global reactions to AI innovations. *Sci Rep.* Sep 15, 2023;13(1):15310. [doi: [10.1038/s41598-023-42227-6](https://doi.org/10.1038/s41598-023-42227-6)] [Medline: [37714915](https://pubmed.ncbi.nlm.nih.gov/37714915/)]
10. Kufel J, Bargieł-Łączek K, Kocot S, et al. What as machine learning, artificial neural networks and deep learning?- examples of practical applications in medicine. *Diagn (Basel).* Aug 3, 2023;13(15):2582. [doi: [10.3390/diagnostics13152582](https://doi.org/10.3390/diagnostics13152582)] [Medline: [37568945](https://pubmed.ncbi.nlm.nih.gov/37568945/)]
11. Mahdavi Ardestani SF, Adibi S, Golshan A, Sadeghian P. Factors influencing the effectiveness of e-learning in healthcare: a fuzzy ANP study. *Healthcare (Basel).* Jul 16, 2023;11(14):2035. [doi: [10.3390/healthcare11142035](https://doi.org/10.3390/healthcare11142035)] [Medline: [37510477](https://pubmed.ncbi.nlm.nih.gov/37510477/)]
12. Habibzadeh F. GPTZero performance in identifying artificial intelligence-generated medical texts: a preliminary study. *J Korean Med Sci.* Sep 25, 2023;38(38):e319. [doi: [10.3346/jkms.2023.38.e319](https://doi.org/10.3346/jkms.2023.38.e319)] [Medline: [37750374](https://pubmed.ncbi.nlm.nih.gov/37750374/)]
13. Gao CA, Howard FM, Markov NS, et al. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *NPJ Digit Med.* Apr 26, 2023;6(1):75. [doi: [10.1038/s41746-023-00819-6](https://doi.org/10.1038/s41746-023-00819-6)] [Medline: [37100871](https://pubmed.ncbi.nlm.nih.gov/37100871/)]

14. Cheng SL, Tsai SJ, Bai YM, et al. Comparisons of quality, correctness, and similarity between chatGPT-generated and human-written abstracts for basic research: cross-sectional study. *J Med Internet Res*. Dec 25, 2023;25:e51229. [doi: [10.2196/51229](https://doi.org/10.2196/51229)] [Medline: [38145486](https://pubmed.ncbi.nlm.nih.gov/38145486/)]
15. Bellini V, Semeraro F, Montomoli J, Cascella M, Bignami E. Between human and AI: assessing the reliability of AI text detection tools. *Curr Med Res Opin*. Mar 3, 2024;40(3):353-358. [doi: [10.1080/03007995.2024.2310086](https://doi.org/10.1080/03007995.2024.2310086)]
16. Edwards B. OpenAI discontinues its AI writing detector due to “low rate of accuracy”. *Ars Technica*. 2023. URL: <https://arstechnica.com/information-technology/2023/07/openai-discontinues-its-ai-writing-detector-due-to-low-rate-of-accuracy/> [Accessed 2025-01-09]
17. Guleria A, Krishan K, Sharma V, Kanchan T. ChatGPT: ethical concerns and challenges in academics and research. *J Infect Dev Ctries*. Sep 30, 2023;17(9):1292-1299. [doi: [10.3855/jidc.18738](https://doi.org/10.3855/jidc.18738)] [Medline: [37824352](https://pubmed.ncbi.nlm.nih.gov/37824352/)]
18. Alser M, Waisberg E. Concerns with the usage of chatGPT in academia and medicine: a viewpoint. *Am J Med Open*. Jun 2023;9:100036. [doi: [10.1016/j.ajmo.2023.100036](https://doi.org/10.1016/j.ajmo.2023.100036)] [Medline: [39035060](https://pubmed.ncbi.nlm.nih.gov/39035060/)]
19. Kadayam Guruswami G, Mumtaz S, Gopakumar A, Khan E, Abdullah F, Parahoo SK. Academic integrity perceptions among health-professions’ students: a cross-sectional study in the middle east. *J Acad Ethics*. 2023;21(2):231-249. [doi: [10.1007/s10805-022-09452-6](https://doi.org/10.1007/s10805-022-09452-6)] [Medline: [35815317](https://pubmed.ncbi.nlm.nih.gov/35815317/)]
20. Crotty D. The latest “crisis” - is the research literature overrun with ChatGPT- and LLM-generated articles? The Scholarly Kitchen. 2024. URL: <https://scholarlykitchen.sspnet.org/2024/03/20/the-latest-crisis-is-the-research-literature-overrun-with-chatgpt-and-llm-generated-articles/> [Accessed 2025-01-09]
21. Mehmet F. ChatGPT: bullshit spewer or the end of traditional assessments in higher education? *JALT*. 2023;6(1). URL: <https://journals.sfu.ca/jalt/index.php/jalt/issue/view/31> [doi: [10.37074/jalt.2023.6.1.9](https://doi.org/10.37074/jalt.2023.6.1.9)]
22. Kasneci E, Sessler K, Küchemann S, et al. ChatGPT for good? On opportunities and challenges of large language models for education. *Learn Individ Differ*. Apr 2023;103:102274. [doi: [10.1016/j.lindif.2023.102274](https://doi.org/10.1016/j.lindif.2023.102274)]
23. Anderson N, Belavy DL, Perle SM, et al. AI did not write this manuscript, or did it? Can we trick the AI text detector into generated texts? The potential future of ChatGPT and AI in sports & exercise medicine manuscript generation. *BMJ Open Sport Exerc Med*. 2023;9(1):e001568. [doi: [10.1136/bmjsem-2023-001568](https://doi.org/10.1136/bmjsem-2023-001568)] [Medline: [36816423](https://pubmed.ncbi.nlm.nih.gov/36816423/)]
24. Lee H. The rise of chatgpt: exploring its potential in medical education. *Anat Sci Ed*. Jul 2024;17(5):926-931. [doi: [10.1002/ase.2270](https://doi.org/10.1002/ase.2270)]
25. Saenger JA, Hunger J, Boss A, Richter J. Delayed diagnosis of a transient ischemic attack caused by ChatGPT. *Wien Klin Wochenschr*. Apr 2024;136(7-8):236-238. [doi: [10.1007/s00508-024-02329-1](https://doi.org/10.1007/s00508-024-02329-1)] [Medline: [38305909](https://pubmed.ncbi.nlm.nih.gov/38305909/)]
26. Khowaja SA, Khuwaja P, Dev K, Wang W, Nkenyereye L, Fortino G. ChatGPT needs SPADE (sustainability, privacy, digital divide, and ethics) evaluation: a review. *TechRxiv*. Preprint posted online on Nov 9, 2023. [doi: [10.36227/techrxiv.22619932.v4](https://doi.org/10.36227/techrxiv.22619932.v4)]
27. Nguyen T. ChatGPT in medical education: a precursor for automation bias? *JMIR Med Educ*. Jan 17, 2024;10:e50174. [doi: [10.2196/50174](https://doi.org/10.2196/50174)] [Medline: [38231545](https://pubmed.ncbi.nlm.nih.gov/38231545/)]
28. Goodman RS, Patrinely JR, Stone CA Jr, et al. Accuracy and reliability of chatbot responses to physician questions. *JAMA Netw Open*. Oct 2, 2023;6(10):e2336483. [doi: [10.1001/jamanetworkopen.2023.36483](https://doi.org/10.1001/jamanetworkopen.2023.36483)] [Medline: [37782499](https://pubmed.ncbi.nlm.nih.gov/37782499/)]
29. Kisekka V, Giboney JS. The effectiveness of health care information technologies: evaluation of trust, security beliefs, and privacy as determinants of health care outcomes. *J Med Internet Res*. Apr 11, 2018;20(4):e107. [doi: [10.2196/jmir.9014](https://doi.org/10.2196/jmir.9014)] [Medline: [29643052](https://pubmed.ncbi.nlm.nih.gov/29643052/)]
30. Ngo E, Patel N, Chandrasekaran K, Tajik AJ, Paterick TE. The importance of the medical record: a critical professional responsibility. *J Med Pract Manage*. 2016;31(5):305-308. [Medline: [27249883](https://pubmed.ncbi.nlm.nih.gov/27249883/)]
31. Gallifant J, Fiske A, Levites Strelakova YA, et al. Peer review of GPT-4 technical report and systems card. *PLOS Dig Health*. Jan 2024;3(1):e0000417. [doi: [10.1371/journal.pdig.0000417](https://doi.org/10.1371/journal.pdig.0000417)] [Medline: [38236824](https://pubmed.ncbi.nlm.nih.gov/38236824/)]
32. Goddard J. Hallucinations in ChatGPT: a cautionary tale for biomedical researchers. *Am J Med*. Nov 2023;136(11):1059-1060. [doi: [10.1016/j.amjmed.2023.06.012](https://doi.org/10.1016/j.amjmed.2023.06.012)] [Medline: [37369274](https://pubmed.ncbi.nlm.nih.gov/37369274/)]
33. Abd-Alrazaq A, AlSaad R, Alhuwail D, et al. Large language models in medical education: opportunities, challenges, and future directions. *JMIR Med Educ*. Jun 1, 2023;9:e48291. [doi: [10.2196/48291](https://doi.org/10.2196/48291)] [Medline: [37261894](https://pubmed.ncbi.nlm.nih.gov/37261894/)]
34. Waisberg E, Ong J, Masalkhi M, Lee AG. Large language model (LLM)-driven chatbots for neuro-ophthalmic medical education. *Eye (Lond)*. Mar 2024;38(4):639-641. [doi: [10.1038/s41433-023-02759-7](https://doi.org/10.1038/s41433-023-02759-7)] [Medline: [37749374](https://pubmed.ncbi.nlm.nih.gov/37749374/)]
35. Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus*. Feb 2023;15(2):e35179. [doi: [10.7759/cureus.35179](https://doi.org/10.7759/cureus.35179)] [Medline: [36811129](https://pubmed.ncbi.nlm.nih.gov/36811129/)]
36. Zielinski C, Winker M, Aggarwal R, et al. Chatbots, ChatGPT, and scholarly manuscripts: WAME recommendations on ChatGPT and chatbots in relation to scholarly publications. *Open Access Maced J Med Sci*. Jul 28, 2023;11(A):83-86. [doi: [10.3889/oamjms.2023.11502](https://doi.org/10.3889/oamjms.2023.11502)]

37. Wei Q, Yao Z, Cui Y, Wei B, Jin Z, Xu X. Evaluation of ChatGPT-generated medical responses: a systematic review and meta-analysis. *J Biomed Inform.* Mar 2024;151:104620. [doi: [10.1016/j.jbi.2024.104620](https://doi.org/10.1016/j.jbi.2024.104620)]
38. Yeo YH, Samaan JS, Ng WH, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. *Clin Mol Hepatol.* Jul 2023;29(3):721-732. [doi: [10.3350/cmh.2023.0089](https://doi.org/10.3350/cmh.2023.0089)]
39. Walker HL, Ghani S, Kuemmerli C, et al. Reliability of medical information provided by ChatGPT: assessment against clinical guidelines and patient information quality instrument. *J Med Internet Res.* Jun 30, 2023;25:e47479. [doi: [10.2196/47479](https://doi.org/10.2196/47479)] [Medline: [37389908](https://pubmed.ncbi.nlm.nih.gov/37389908/)]
40. Ali MJ. ChatGPT and lacrimal drainage disorders: performance and scope of improvement. *Ophthalmic Plast Reconstr Surg.* 2023;39(3):221-225. [doi: [10.1097/IOP.0000000000002418](https://doi.org/10.1097/IOP.0000000000002418)] [Medline: [37166289](https://pubmed.ncbi.nlm.nih.gov/37166289/)]
41. Van Bulck L, Moons P. What if your patient switches from Dr. Google to Dr. ChatGPT? A vignette-based survey of the trustworthiness, value, and danger of ChatGPT-generated responses to health questions. *Eur J Cardiovasc Nurs.* Jan 12, 2024;23(1):95-98. [doi: [10.1093/eurjcn/zvad038](https://doi.org/10.1093/eurjcn/zvad038)]
42. Cankurtaran RE, Polat YH, Aydemir NG, Umay E, Yurekli OT. Reliability and usefulness of ChatGPT for inflammatory bowel diseases: an analysis for patients and healthcare professionals. *Cureus.* Oct 2023;15(10):e46736. [doi: [10.7759/cureus.46736](https://doi.org/10.7759/cureus.46736)] [Medline: [38022227](https://pubmed.ncbi.nlm.nih.gov/38022227/)]
43. Athaluri SA, Manthena SV, Kesapragada V, Yarlagadda V, Dave T, Duddumpudi RTS. Exploring the boundaries of reality: investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. *Cureus.* Apr 2023;15(4):e37432. [doi: [10.7759/cureus.37432](https://doi.org/10.7759/cureus.37432)] [Medline: [37182055](https://pubmed.ncbi.nlm.nih.gov/37182055/)]
44. Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J. Ethical considerations of using ChatGPT in health care. *J Med Internet Res.* Aug 11, 2023;25:e48009. [doi: [10.2196/48009](https://doi.org/10.2196/48009)] [Medline: [37566454](https://pubmed.ncbi.nlm.nih.gov/37566454/)]
45. ChatGPT plugins. OpenAI. URL: <https://openai.com/index/chatgpt-plugins/> [Accessed 2025-01-09]
46. Chen H, Hailey D, Wang N, Yu P. A review of data quality assessment methods for public health information systems. *Int J Environ Res Public Health.* May 14, 2014;11(5):5170-5207. [doi: [10.3390/ijerph110505170](https://doi.org/10.3390/ijerph110505170)] [Medline: [24830450](https://pubmed.ncbi.nlm.nih.gov/24830450/)]
47. Hatia A, Doldo T, Parrini S, et al. Accuracy and completeness of ChatGPT-generated information on interceptive orthodontics: a multicenter collaborative study. *J Clin Med.* Jan 27, 2024;13(3):735. [doi: [10.3390/jcm13030735](https://doi.org/10.3390/jcm13030735)] [Medline: [38337430](https://pubmed.ncbi.nlm.nih.gov/38337430/)]
48. Lakdawala N, Channa L, Gronbeck C, et al. Assessing the accuracy and comprehensiveness of ChatGPT in offering clinical guidance for atopic dermatitis and acne vulgaris. *JMIR Dermatol.* Nov 14, 2023;6:e50409. [doi: [10.2196/50409](https://doi.org/10.2196/50409)] [Medline: [37962920](https://pubmed.ncbi.nlm.nih.gov/37962920/)]
49. Homolak J. Opportunities and risks of ChatGPT in medicine, science, and academic publishing: a modern Promethean dilemma. *Croat Med J.* Feb 2023;64(1):1-3. [doi: [10.3325/cmj.2023.64.1](https://doi.org/10.3325/cmj.2023.64.1)]
50. Deng J, Heybati K, Park YJ, Zhou F, Bozzo A. Artificial intelligence in clinical practice: a look at ChatGPT. *Cleve Clin J Med.* Mar 1, 2024;91(3):173-180. [doi: [10.3949/ccjm.91a.23070](https://doi.org/10.3949/ccjm.91a.23070)] [Medline: [38429005](https://pubmed.ncbi.nlm.nih.gov/38429005/)]
51. GPT-4 is openai's most advanced system, producing safer and more useful responses. OpenAI. URL: <https://openai.com/gpt-4> [Accessed 2024-03-06]
52. Google Gemini: what is it, and how does it work? XDA. URL: <https://www.xda-developers.com/google-bard/> [Accessed 2024-05-01]
53. Zhang X, Li S, Hauer B, Shi N, Kondrak G. Don't trust ChatGPT when your question is not in English: a study of multilingual abilities and types of LLMs. Presented at: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; Dec 6-10, 2023; Singapore. [doi: [10.18653/v1/2023.emnlp-main.491](https://doi.org/10.18653/v1/2023.emnlp-main.491)]
54. Lai VD, Ngo NT, Veyseh APB, et al. ChatGPT beyond English: towards a comprehensive evaluation of large language models in multilingual learning. *arXiv.* Preprint posted online on Apr 12, 2023. [doi: [10.48550/arXiv.2304.05613](https://doi.org/10.48550/arXiv.2304.05613)]
55. Wang X, Sanders HM, Liu Y, et al. ChatGPT: promise and challenges for deployment in low- and middle-income countries. *Lancet Reg Health West Pac.* Dec 2023;41:100905. [doi: [10.1016/j.lanwpc.2023.100905](https://doi.org/10.1016/j.lanwpc.2023.100905)] [Medline: [37731897](https://pubmed.ncbi.nlm.nih.gov/37731897/)]
56. Kuzdeuov A, Nurgaliyev S, Varol HA. ChatGPT for visually impaired and blind. *TechRxiv.* Preprint posted online on May 3, 2023. [doi: [10.36227/techrxiv.22047080](https://doi.org/10.36227/techrxiv.22047080)]
57. A social robot connected with ChatGPT to improve cognitive functioning in ASD subjects. *Frontiers.* URL: <https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2023.1232177/full> [Accessed 2025-01-09]
58. Adhikari K, Naik N, Hameed BZ, Raghunath SK, Somani BK. Exploring the ethical, legal, and social implications of ChatGPT in urology. *Curr Urol Rep.* Jan 2024;25(1):1-8. [doi: [10.1007/s11934-023-01185-2](https://doi.org/10.1007/s11934-023-01185-2)] [Medline: [37735339](https://pubmed.ncbi.nlm.nih.gov/37735339/)]
59. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese medical licensing examination: comparison study. *JMIR Med Educ.* Jun 29, 2023;9:e48002. [doi: [10.2196/48002](https://doi.org/10.2196/48002)] [Medline: [37384388](https://pubmed.ncbi.nlm.nih.gov/37384388/)]

60. Meyer A, Riese J, Streichert T. Comparison of the performance of GPT-3.5 and GPT-4 with that of medical students on the written German medical licensing examination: observational study. *JMIR Med Educ.* Feb 8, 2024;10:e50965. [doi: [10.2196/50965](https://doi.org/10.2196/50965)] [Medline: [38329802](https://pubmed.ncbi.nlm.nih.gov/38329802/)]
61. Jang D, Yun TR, Lee CY, Kwon YK, Kim CE. GPT-4 can pass the Korean national licensing examination for Korean medicine doctors. *PLOS Dig Health.* Dec 2023;2(12):e0000416. [doi: [10.1371/journal.pdig.0000416](https://doi.org/10.1371/journal.pdig.0000416)] [Medline: [38100393](https://pubmed.ncbi.nlm.nih.gov/38100393/)]
62. Farhat F, Chaudhry BM, Nadeem M, Sohail SS, Madsen DØ. Evaluating large language models for the national premedical exam in India: comparative analysis of GPT-3.5, GPT-4, and Bard. *JMIR Med Educ.* Feb 21, 2024;10:e51523. [doi: [10.2196/51523](https://doi.org/10.2196/51523)] [Medline: [38381486](https://pubmed.ncbi.nlm.nih.gov/38381486/)]
63. Oh N, Choi GS, Lee WY. ChatGPT goes to the operating room: evaluating GPT-4 performance and its potential in surgical education and training in the era of large language models. *Ann Surg Treat Res.* 2023;104(5):269. [doi: [10.4174/astr.2023.104.5.269](https://doi.org/10.4174/astr.2023.104.5.269)]
64. Nakajima N, Fujimori T, Furuya M, et al. A comparison between GPT-3.5, GPT-4, and GPT-4V: can the large language model (ChatGPT) pass the Japanese board of orthopaedic surgery examination? *Cureus.* Mar 2024;16(3):e56402. [doi: [10.7759/cureus.56402](https://doi.org/10.7759/cureus.56402)] [Medline: [38633935](https://pubmed.ncbi.nlm.nih.gov/38633935/)]
65. Rojas M, Rojas M, Burgess V, Toro-Pérez J, Salehi S. Exploring the performance of ChatGPT versions 3.5, 4, and 4 with vision in the Chilean medical licensing examination: observational study. *JMIR Med Educ.* Apr 29, 2024;10:e55048. [doi: [10.2196/55048](https://doi.org/10.2196/55048)] [Medline: [38686550](https://pubmed.ncbi.nlm.nih.gov/38686550/)]
66. Moshirfar M, Altaf AW, Stoakes IM, Tuttle JJ, Hoopes PC. Artificial intelligence in ophthalmology: a comparative analysis of GPT-3.5, GPT-4, and human expertise in answering StatPearls questions. *Cureus.* Jun 2023;15(6):e40822. [doi: [10.7759/cureus.40822](https://doi.org/10.7759/cureus.40822)] [Medline: [37485215](https://pubmed.ncbi.nlm.nih.gov/37485215/)]
67. Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. *Neurosurgery.* Dec 1, 2023;93(6):1353-1365. [doi: [10.1227/neu.0000000000002632](https://doi.org/10.1227/neu.0000000000002632)] [Medline: [37581444](https://pubmed.ncbi.nlm.nih.gov/37581444/)]
68. Brin D, Sorin V, Vaid A, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. *Sci Rep.* Oct 1, 2023;13(1):16492. [doi: [10.1038/s41598-023-43436-9](https://doi.org/10.1038/s41598-023-43436-9)] [Medline: [37779171](https://pubmed.ncbi.nlm.nih.gov/37779171/)]
69. Gill GS, Blair J, Litinsky S. Evaluating the performance of ChatGPT 3.5 and 4.0 on StatPearls oculoplastic surgery text- and image-based exam questions. *Cureus.* Nov 2024;16(11):e73812. [doi: [10.7759/cureus.73812](https://doi.org/10.7759/cureus.73812)] [Medline: [39691123](https://pubmed.ncbi.nlm.nih.gov/39691123/)]
70. Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT, GPT-4, and Google Bard on a neurosurgery oral boards preparation question bank. *medRxiv.* Preprint posted online on 2023. [doi: [10.1101/2023.04.06.23288265](https://doi.org/10.1101/2023.04.06.23288265)]
71. Taloni A, Borselli M, Scarsi V, et al. Comparative performance of humans versus GPT-4.0 and GPT-3.5 in the self-assessment program of American Academy of Ophthalmology. *Sci Rep.* Oct 29, 2023;13(1):18562. [doi: [10.1038/s41598-023-45837-2](https://doi.org/10.1038/s41598-023-45837-2)] [Medline: [37899405](https://pubmed.ncbi.nlm.nih.gov/37899405/)]
72. Lee GU, Hong DY, Kim SY, et al. Comparison of the problem-solving performance of ChatGPT-3.5, ChatGPT-4, Bing Chat, and Bard for the Korean emergency medicine board examination question bank. *Medicine (Abingdon).* 2024;103(9):e37325. [doi: [10.1097/MD.00000000000037325](https://doi.org/10.1097/MD.00000000000037325)]
73. Al-Ashwal FY, Zawiah M, Gharaibeh L, Abu-Farha R, Bitar AN. Evaluating the sensitivity, specificity, and accuracy of ChatGPT-3.5, ChatGPT-4, Bing AI, and Bard against conventional drug-drug interactions clinical tools. *Drug Healthc Patient Saf.* 2023;15:137-147. [doi: [10.2147/DHPS.S425858](https://doi.org/10.2147/DHPS.S425858)] [Medline: [37750052](https://pubmed.ncbi.nlm.nih.gov/37750052/)]
74. Agharia S, Szatkowski J, Fraval A, Stevens J, Zhou Y. The ability of artificial intelligence tools to formulate orthopaedic clinical decisions in comparison to human clinicians: An analysis of ChatGPT 3.5, ChatGPT 4, and Bard. *J Orthop.* Apr 2024;50:1-7. [doi: [10.1016/j.jor.2023.11.063](https://doi.org/10.1016/j.jor.2023.11.063)] [Medline: [38148925](https://pubmed.ncbi.nlm.nih.gov/38148925/)]
75. Lechien JR, Briganti G, Vaira LA. Accuracy of ChatGPT-3.5 and -4 in providing scientific references in otolaryngology-head and neck surgery. *Eur Arch Otorhinolaryngol.* Apr 2024;281(4):2159-2165. [doi: [10.1007/s00405-023-08441-8](https://doi.org/10.1007/s00405-023-08441-8)]
76. Balasanjeevi G, Surapaneni KM. Comparison of ChatGPT version 3.5 & 4 for utility in respiratory medicine education using clinical case scenarios. *Respir Med Res.* Jun 2024;85:101091. [doi: [10.1016/j.resmer.2024.101091](https://doi.org/10.1016/j.resmer.2024.101091)] [Medline: [38657295](https://pubmed.ncbi.nlm.nih.gov/38657295/)]
77. Liang R, Zhao A, Peng L, et al. Enhanced artificial intelligence strategies in renal oncology: iterative optimization and comparative analysis of GPT 3.5 versus 4.0. *Ann Surg Oncol.* Jun 2024;31(6):3887-3893. [doi: [10.1245/s10434-024-15107-0](https://doi.org/10.1245/s10434-024-15107-0)] [Medline: [38472675](https://pubmed.ncbi.nlm.nih.gov/38472675/)]
78. Momenai B, Wakabayashi T, Shahlaee A, et al. Assessing ChatGPT-3.5 versus ChatGPT-4 performance in surgical treatment of retinal diseases: a comparative study. *Ophthalmic Surg Lasers Imaging Retina.* Aug 2024;55(8):481-482. [doi: [10.3928/23258160-20240227-02](https://doi.org/10.3928/23258160-20240227-02)] [Medline: [38531015](https://pubmed.ncbi.nlm.nih.gov/38531015/)]
79. Samaan JS, Rajeev N, Ng WH, et al. ChatGPT as a source of information for bariatric surgery patients: a comparative analysis of accuracy and comprehensiveness between GPT-4 and GPT-3.5. *Obes Surg.* May 2024;34(5):1987-1989. [doi: [10.1007/s11695-024-07212-6](https://doi.org/10.1007/s11695-024-07212-6)] [Medline: [38564173](https://pubmed.ncbi.nlm.nih.gov/38564173/)]

80. Singh OP. Artificial intelligence in the era of ChatGPT - Opportunities and challenges in mental health care. *Indian J Psychiatry*. Mar 2023;65(3):297-298. [doi: [10.4103/indianjpsychiatry.indianjpsychiatry_112_23](https://doi.org/10.4103/indianjpsychiatry.indianjpsychiatry_112_23)] [Medline: [37204980](https://pubmed.ncbi.nlm.nih.gov/37204980/)]
81. Adams-Prassl J, Binns R, Kelly-Lyth A. Directly discriminatory algorithms. *Mod Law Rev*. Jan 2023;86(1):144-175. [doi: [10.1111/1468-2230.12759](https://doi.org/10.1111/1468-2230.12759)] [Medline: [37065788](https://pubmed.ncbi.nlm.nih.gov/37065788/)]
82. Hoffman S, Podgurski A. Artificial intelligence and discrimination in health care. *Yale J Health Policy Law Ethics*. 2020;19(3). URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3747737 [Accessed 2025-01-09]
83. Amin KS, Forman HP, Davis MA. Even with ChatGPT, race matters. *Clin Imaging*. May 2024;109:110113. [doi: [10.1016/j.clinimag.2024.110113](https://doi.org/10.1016/j.clinimag.2024.110113)]
84. Hamed S, Bradby H, Ahlberg BM, Thapar-Björkert S. Racism in healthcare: a scoping review. *BMC Public Health*. May 16, 2022;22(1):988. [doi: [10.1186/s12889-022-13122-y](https://doi.org/10.1186/s12889-022-13122-y)] [Medline: [35578322](https://pubmed.ncbi.nlm.nih.gov/35578322/)]
85. Omiye JA, Lester JC, Spichak S, Rotemberg V, Daneshjou R. Large language models propagate race-based medicine. *NPJ Digit Med*. Oct 20, 2023;6(1):195. [doi: [10.1038/s41746-023-00939-z](https://doi.org/10.1038/s41746-023-00939-z)] [Medline: [37864012](https://pubmed.ncbi.nlm.nih.gov/37864012/)]
86. Hoffman KM, Trawalter S, Axt JR, Oliver MN. Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites. *Proc Natl Acad Sci U S A*. Apr 19, 2016;113(16):4296-4301. [doi: [10.1073/pnas.1516047113](https://doi.org/10.1073/pnas.1516047113)] [Medline: [27044069](https://pubmed.ncbi.nlm.nih.gov/27044069/)]
87. Tsai JW, Cerdeña JP, Goedel WC, et al. Evaluating the impact and rationale of race-specific estimations of kidney function: estimations from U.S. NHANES, 2015-2018. *E Clin Med*. Dec 2021;42:101197. [doi: [10.1016/j.eclinm.2021.101197](https://doi.org/10.1016/j.eclinm.2021.101197)] [Medline: [34849475](https://pubmed.ncbi.nlm.nih.gov/34849475/)]
88. Zack T, Lehman E, Suzgun M, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. *Lancet Dig Health*. Jan 2024;6(1):e12-e22. [doi: [10.1016/S2589-7500\(23\)00225-X](https://doi.org/10.1016/S2589-7500(23)00225-X)]
89. Yang Y, Liu X, Jin Q, Huang F, Lu Z. Unmasking and quantifying racial bias of large language models in medical report generation. *Commun Med (Lond)*. Sep 10, 2024;4(1):176. [doi: [10.1038/s43856-024-00601-z](https://doi.org/10.1038/s43856-024-00601-z)] [Medline: [39256622](https://pubmed.ncbi.nlm.nih.gov/39256622/)]
90. Koskenvuori J, Stolt M, Suhonen R, Leino-Kilpi H. Healthcare professionals' ethical competence: a scoping review. *Nurs Open*. Jan 2019;6(1):5-17. [doi: [10.1002/nop2.173](https://doi.org/10.1002/nop2.173)] [Medline: [30534390](https://pubmed.ncbi.nlm.nih.gov/30534390/)]
91. Iyalomhe GBS. Medical ethics and ethical dilemmas. *Niger J Med*. 2009;18(1):8-16. [Medline: [19485140](https://pubmed.ncbi.nlm.nih.gov/19485140/)]
92. Andersson H, Svensson A, Frank C, Rantala A, Holmberg M, Bremer A. Ethics education to support ethical competence learning in healthcare: an integrative systematic review. *BMC Med Ethics*. Mar 19, 2022;23(1):29. [doi: [10.1186/s12910-022-00766-z](https://doi.org/10.1186/s12910-022-00766-z)] [Medline: [35305627](https://pubmed.ncbi.nlm.nih.gov/35305627/)]
93. Krügel S, Ostermaier A, Uhl M. ChatGPT's inconsistent moral advice influences users' judgment. *Sci Rep*. Apr 6, 2023;13(1):4569. [doi: [10.1038/s41598-023-31341-0](https://doi.org/10.1038/s41598-023-31341-0)] [Medline: [37024502](https://pubmed.ncbi.nlm.nih.gov/37024502/)]
94. Balas M, Wadden JJ, Hébert PC, et al. Exploring the potential utility of AI large language models for medical ethics: an expert panel evaluation of GPT-4. *J Med Ethics*. Feb 2024;50(2):90-96. [doi: [10.1136/jme-2023-109549](https://doi.org/10.1136/jme-2023-109549)]
95. Franco D'Souza R, Mathew M, Louis Palatty P, Surapaneni KM. Teaching humanism with humanoid: evaluating the potential of ChatGPT-4 as a pedagogical tool in bioethics education using validated clinical case vignettes. *Int J Ethics Educ*. Apr 30, 2024. [doi: [10.1007/s40889-024-00190-4](https://doi.org/10.1007/s40889-024-00190-4)]
96. Chen J, Cadiante A, Kasselmann LJ, Pilkington B. Assessing the performance of ChatGPT in bioethics: a large language model's moral compass in medicine. *J Med Ethics*. Feb 2024;50(2):97-101. [doi: [10.1136/jme-2023-109366](https://doi.org/10.1136/jme-2023-109366)]
97. Singh G, Deng F, Ahn S. Illiterate DALL-E learns to compose. *arXiv*. Preprint posted online on Oct 17, 2021. [doi: [10.48550/arXiv.2110.11405](https://doi.org/10.48550/arXiv.2110.11405)]
98. Adams LC, Busch F, Truhn D, Makowski MR, Aerts H, Bresssem KK. What does DALL-E 2 know about radiology? *J Med Internet Res*. Mar 16, 2023;25:e43110. [doi: [10.2196/43110](https://doi.org/10.2196/43110)] [Medline: [36927634](https://pubmed.ncbi.nlm.nih.gov/36927634/)]
99. Zhu L, Mou W, Wu K, Zhang J, Luo P. Can DALL-E 3 reliably generate 12-Lead ECGs and teaching illustrations? *Cureus*. Jan 2024;16(1):e52748. [doi: [10.7759/cureus.52748](https://doi.org/10.7759/cureus.52748)] [Medline: [38384621](https://pubmed.ncbi.nlm.nih.gov/38384621/)]
100. Cheraghlou S. Evaluating dermatologic domain knowledge in DALL-E 2 and potential applications for dermatology-specific algorithms. *Int J Dermatol*. Oct 2023;62(10):e521-e523. [doi: [10.1111/ijd.16683](https://doi.org/10.1111/ijd.16683)] [Medline: [37057623](https://pubmed.ncbi.nlm.nih.gov/37057623/)]
101. DALL-E 2. OpenAI. URL: <https://openai.com/index/dall-e-2> [Accessed 2025-01-09]
102. Why is chatgpt bad at math? Baeldung. 2023. URL: <https://www.baeldung.com/cs/chatgpt-math-problems> [Accessed 2025-01-09]
103. Ignjatović A, Stevanović L. Efficacy and limitations of ChatGPT as a biostatistical problem-solving tool in medical education in Serbia: a descriptive study. *J Educ Eval Health Prof*. 2023;20:28. [doi: [10.3352/jeehp.2023.20.28](https://doi.org/10.3352/jeehp.2023.20.28)] [Medline: [37840252](https://pubmed.ncbi.nlm.nih.gov/37840252/)]
104. Ordak M. ChatGPT's skills in statistical analysis using the example of allergology: do we have reason for concern? *Healthc (Basel)*. Sep 15, 2023;11(18):2554. [doi: [10.3390/healthcare11182554](https://doi.org/10.3390/healthcare11182554)] [Medline: [37761751](https://pubmed.ncbi.nlm.nih.gov/37761751/)]

105. Huang Y, Wu R, He J, Xiang Y. Evaluating ChatGPT-4.0's data analytic proficiency in epidemiological studies: a comparative analysis with SAS, SPSS, and R. *J Glob Health*. Mar 29, 2024;14:04070. [doi: [10.7189/jogh.14.04070](https://doi.org/10.7189/jogh.14.04070)] [Medline: [38547497](https://pubmed.ncbi.nlm.nih.gov/38547497/)]
106. McKay F, Williams BJ, Prestwich G, Bansal D, Treanor D, Hallowell N. Artificial intelligence and medical research databases: ethical review by data access committees. *BMC Med Ethics*. Jul 8, 2023;24(1):49. [doi: [10.1186/s12910-023-00927-8](https://doi.org/10.1186/s12910-023-00927-8)] [Medline: [37422629](https://pubmed.ncbi.nlm.nih.gov/37422629/)]
107. Rothstein MA. Is deidentification sufficient to protect health privacy in research? *Am J Bioeth*. Sep 2010;10(9):3-11. [doi: [10.1080/15265161.2010.494215](https://doi.org/10.1080/15265161.2010.494215)] [Medline: [20818545](https://pubmed.ncbi.nlm.nih.gov/20818545/)]
108. Baumgartner C. The potential impact of ChatGPT in clinical and translational medicine. *Clin Transl Med*. Mar 2023;13(3):e1206. [doi: [10.1002/ctm2.1206](https://doi.org/10.1002/ctm2.1206)] [Medline: [36854881](https://pubmed.ncbi.nlm.nih.gov/36854881/)]
109. Terms of use. OpenAI. URL: <https://openai.com/policies/terms-of-use> [Accessed 2024-05-03]
110. Cohen IG. What should ChatGPT mean for bioethics? *Am J Bioeth*. Oct 2023;23(10):8-16. [doi: [10.1080/15265161.2023.2233357](https://doi.org/10.1080/15265161.2023.2233357)] [Medline: [37440696](https://pubmed.ncbi.nlm.nih.gov/37440696/)]
111. Else H. Abstracts written by ChatGPT fool scientists. *Nature New Biol*. Jan 2023;613(7944):423. [doi: [10.1038/d41586-023-00056-7](https://doi.org/10.1038/d41586-023-00056-7)] [Medline: [36635510](https://pubmed.ncbi.nlm.nih.gov/36635510/)]
112. Mai KT, Bray S, Davies T, Griffin LD. Warning: humans cannot reliably detect speech deepfakes. *PLoS ONE*. 2023;18(8):e0285333. [doi: [10.1371/journal.pone.0285333](https://doi.org/10.1371/journal.pone.0285333)] [Medline: [37531336](https://pubmed.ncbi.nlm.nih.gov/37531336/)]
113. Hoehe MR, Thibaut F. Going digital: how technology use may influence human brains and behavior. *Dialogues Clin Neurosci*. Jun 2020;22(2):93-97. [doi: [10.31887/DCNS.2020.22.2/mhoehe](https://doi.org/10.31887/DCNS.2020.22.2/mhoehe)] [Medline: [32699509](https://pubmed.ncbi.nlm.nih.gov/32699509/)]

Abbreviations

AI: artificial intelligence

GPT: generative pretrained transformer

LLM: large language model

Edited by Amaryllis Mavragani; peer-reviewed by Ela Vashishtha, Shayan Fazeli; submitted 28.07.2023; final revised version received 31.08.2024; accepted 03.09.2024; published 16.01.2025

Please cite as:

Kim J, Vajravelu BN

Assessing the Current Limitations of Large Language Models in Advancing Health Care Education

JMIR Form Res 2025;9:e51319

URL: <https://formative.jmir.org/2025/1/e51319>

doi: [10.2196/51319](https://doi.org/10.2196/51319)

© JaeYong Kim, Bathri Narayan Vajravelu. Originally published in JMIR Formative Research (<https://formative.jmir.org>), 16.01.2025. 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 Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on <https://formative.jmir.org>, as well as this copyright and license information must be included.