npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-024-01251-0

Results and implications for generative AI in a large introductory biomedical and health informatics course

Check for updates

William Hersh ✉ & Kate Fultz Hollis

Generative artificial intelligence (AI) systems have performed well at many biomedical tasks, but few studies have assessed their performance directly compared to students in higher-education courses. We compared student knowledge-assessment scores with prompting of 6 large-language model (LLM) systems as they would be used by typical students in a large online introductory course in biomedical and health informatics that is taken by graduate, continuing education, and medical students. The state-of-the-art LLM systems were prompted to answer multiple-choice questions (MCQs) and final exam questions. We compared the scores for 139 students (30 graduate students, 85 continuing education students, and 24 medical students) to the LLM systems. All of the LLMs scored between the 50th and 75th percentiles of students for MCQ and final exam questions. The performance of LLMs raises questions about student assessment in higher education, especially in courses that are knowledge-based and online.

Generative artificial intelligence (AI), driven by large language models (LLMs), has had a profound impact in all scientific disciplines. The impacts in biomedicine have spanned across clinical practice, research, and education[1]. In education, LLMs have been shown to score well above passing levels on medical board exams[2–4], although until recently, none has compared scores directly with trainee test-takers on actual tests[5]. LLMs have also been found to perform comparably well with students and others on objective structured clinical examinations[6], answering general-domain clinical questions[7,8], and solving clinical cases[9–13]. They have also been shown to engage in conversational diagnostic dialogue[14] as well as exhibit clinical reasoning comparable to physicians[15]. LLMs have had comparable strong impact in education in fields beyond biomedicine, such as business[16], computer science[17–19], law[20], and data science[21].

The successes of LLMs raise concerns about the future of student learning and assessment, particularly in the higher-education setting. LLMs may be good at providing answers, but they do not necessarily steer users (or students) to the original sources of knowledge nor assess their trustworthiness[22,23]. Another issue is the general tendency of LLMs to hallucinate or otherwise confabulate with stated confidence, potentially misleading students[24]. Others note that that LLMs might give students easy answers to assessments and undermine their learning and development of competence[25–28]. Among the concerns for LLMs are that users find their output competent, trustworthy, clear, and engaging, which may not be warranted[29].

This study aimed to compare how LLMs perform on the assessments in one of the most widely taken online introductory courses in the field of biomedical and health informatics, a course taught at Oregon Health & Science University (OHSU) by one of the authors (WRH) for nearly three decades. The other author (KFH) has been a teaching assistant (TA) in the course for over the last decade. The course is offered to three different audiences using identical curricular materials and assessments:

- Graduate students (BMI 510/610)—this course has been offered as part of what is now the health and clinical informatics (HCIN) major in the OHSU Biomedical Informatics Graduate Program. In addition to students in the HCIN major, students in other graduate programs (e.g., public health, nursing, biomedical basic science, etc.) can take this course as an elective in their programs.
- Continuing education (AMIA 10 × 10)—starting in 2005, this course is known as 10×10 ("ten by ten")[30,31].
- Medical students (MINF 705B/709 A)—beginning early in the COVID-19 pandemic, when medical education had to rapidly pivot to use of virtual learning, this course was offered as an elective for medical students and has continued due to student interest.

All offerings of the course are online. The major curricular activity is voice-over-PowerPoint lectures, with about three hours of lecture for each of the 10 units of the course. Additional readings using a textbook are optional. Students participate in threaded discussion in OHSU's instance of the open-

Department of Medical Informatics & Clinical Epidemiology, School of Medicine, Oregon Health & Science University, 3181 SW Sam Jackson Park Rd. BICC, Portland, OR, USA. ✉e-mail: hersh@ohsu.edu

Table 1 | Biomedical and health informatics introductory course outline units with titles

Unit	Title
1	Overview of Field and Problems Motivating It
2	Computing Concepts for Biomedical and Health Informatics
3	Electronic and Personal Health Records (EHR, PHR)
4	Standards and Interoperability
5	Data Science and Artificial Intelligence
6	Advanced Use of the EHR
7	EHR Implementation, Security, and Evaluation
8	Information Retrieval (Search)
9	Research Informatics
10	Other Areas of Informatics

(Full outline in Supplementary Note 1).

source Sakai learning management system (LMS) and are assessed with multiple-choice questions (MCQs), a final exam, and (for some—see below) a course paper.

On an academic quarter system, OHSU offers BMI 510/610 as a 10-unit course, with units released weekly. Because the AMIA 10 × 10 course is a continuing education course, the 10 units are decompressed and offered over 16 weeks. The medical student version of the course is offered as a two-week block (705B) or over an academic quarter (709 A). From 1996 through the winter quarter of 2024, 1683 students had completed BMI 510/610. From its inception in 2005 through the latest offering ending in early 2024, 3260 individuals had completed the 10×10 course.

Student learning in this course is assessed by up to three activities, depending on the audience:

- MCQs: Each of the 10 units has an assessment of 10 questions per unit that is required for all students.
- Final examination: The exam is required in BMI 510/610; optional in AMIA 10 × 10 for those wanting to obtain academic credit, usually to pursue further study in the field; and not required in MINF 705 A. Students are instructed to provide short answers of one sentence or less on the 33-question exam. The exam has historically been open-book so that students can focus on applying material and not memorizing it. As such, test-takers can consult materials on the LMS and the Internet, although are forbidden from contacting people.
- Course project: A term paper of 10-15 pages is required for BMI 510/610, while a three-page paper is required for AMIA 10 × 10, and none is required for MINF 705 A/709B.

Overall student grading for each course is as follows:

- BMI 510/610 is graded on a letter-grade scale. The final grade is weighted for the MCQs (30%), final exam (30%), student paper (30%), and class participation (10%).
- AMIA 10 × 10 is a continuing-education course and graded on a pass-fail basis. Students completing the course can optionally take the BMI 510/610 final exam to get academic credit for the course, and a letter grade is assigned based on the final exam grade.
- MINF 705B/709 A is graded (as with all OHSU medical school courses) on a pass-fail basis. Students are required to obtain an average of 70% across all of the MCQs and are not required to take the final exam or write a course paper.

The content of the course is updated annually and aims to reflect the latest research findings, operational best practices, government programs and regulations, and future directions for the field. The goal of the course is to provide a detailed overview of biomedical and health informatics to those who will work at the interface of healthcare and information technology (IT). The course also aims to provide an entry point for those wishing further study (and/or career development) in the field. It provides a broad understanding of the field from the vantage point of those who implement, lead, and develop IT solutions for improving health, healthcare, public health, and biomedical research. The annual updating is undertaken at the beginning of each calendar year, with the course materials rolled out in courses starting in the spring. An outline of the course content is listed in Table 1, with more detail provided in Supplementary Note 1.

In this study, we compared the knowledge-assessment results of students with those obtained by prompting several commercial LLMs and one open-source LLM as they would likely be used by higher-education students, i.e., through their interactive Web interfaces. The goal of the study was to assess how well these LLMs performed in a highly subscribed introductory course in biomedical and health informatics compared to realistic use by students Table 2.

Results

The 2023 version of the course was offered between Spring 2023 and Winter 2024. With the 2023 content, the course was completed by a total of 139 students, with 30 graduate students (BMI 510/610), 85 continuing students (AMIA 10 × 10), and 24 medical students (MINF 705 A/709B). The MCQs were answered by all students completing all courses, while all 30 BMI 510/610 students completed the final exam and 21 of 85 students opted to take the final exam in AMIA 10 × 10. The minimum, 25th quartile, median, 75th quartile, and maximum score are shown for MCQs on each of the unit assessments and the final exam for each student group and all groups combined in Table 1.

The output from the LLM prompts of the MCQs and final exam was graded by KFH and is shown in Table 3. ChatGPT Plus and CoPilot-Bing Precise tied for the highest average score on the MCQs, followed closely by Gemini Pro, Llama 3.1 405B, Mistral-Large, and Claude 3 Opus. On the final exam, Gemini Pro and Claude 3 Opus scored highest, followed by Llama 3.1 405B, CoPilot-Bing Precise, Mistral-Large, and ChatGPT Plus. Giving equal weighting to the MCQ average and the final exam, Gemini Pro scored the best overall. Figure 1 summarizes some key results, namely the MCQ averages and final exam results for students at the 25th, 50th (median), and 75th quartile of performance, along with Gemini Pro. Gemini Pro scored above the 75th percentile on 3 unit quizzes, equal to the 75th percentile on 1 unit quiz, and below the 75th percentile on 6 unit quizzes. Gemini Pro scored above the 50th percentile on 4 unit quizzes, equal to the 50th percentile on 4 unit quizzes, and below the 50th percentile on 2 unit quizzes. Gemini Pro scored above the 75th percentile on the final exam.

The stopwatch times taken for each prompt for each LLM are shown in Table 4. Although there were substantial time differences among the LLMs, the time taken for all LLMs was minimal compared to the time taken by students. Although we have no data on time taken to complete MCQs, students are given up to 4 hours to complete the final exam, and the average time taken was 162 minutes (range 34–240). An observation of timing the LLM output was that it was most related to the amount of text each LLM printed to the screen, with some LLMs giving just answers and others providing text explanations of longer length and taking more time to display the text to the browser window.

The distribution of correct and incorrect answers for the LLMs on the final exam is shown in Fig. 2. Every LLM gave wrong answers on questions 19 and 23, the latter of which required students to calculate a Boolean expression. All LLMs answered 23 of the 33 questions correctly.

Discussion

This is, to our knowledge, the first assessment of LLMs based on a course in the biomedical domain where performance was compared with actual student results. In addition, the student assessment data comes from relatively large numbers of learners in three different types of educational programs—graduate, continuing education, and medical student.

All the LLMs performed well on course materials, with Gemini Pro performing best and Llama 3.1 405B, Claude 3 Opus and CoPilot Bing-

Table 2 | Minimum, 25th quartile, median, 75th quartile, maximum, and average scores for MCQs on each unit assessment and the final exam for each student group and all groups

Students	Unit 1	Unit 2	Unit 3	Unit 4	Unit 5	Unit 6	Unit 7	Unit 8	Unit 9	Unit 10	MCQ Average	Final Exam	MCQ+Final Combined
	Graduate students (BMI 510/610), *n* = 30												
Min	70	50	60	50	70	50	60	70	30	70	58	58	116
25th	82.5	70	70	70	82.5	80	80	80	60	90	76.5	76	152.5
Median	90	80	80	70	90	80	90	90	70	90	83	85	168
75th	100	90	90	80	100	90	100	100	77.5	100	92.8	91	183.8
Max	100	100	100	100	100	100	100	100	100	100	100	97	197
Average	90.3	80.3	80.3	74.0	90.4	84.3	88.0	89.3	67.3	92.3	83.7	83.7	167.4
	Continuing education students (AMIA 10 × 10), *n* = 85 for MCQs, *n* = 21 for final exam												
Min	40	40	30	20	20	20	40	20	20	50	30	43	73
25th	80	70	70	60	70	70	80	70	50	80	70	73	143
Median	80	80	80	70	80	80	90	80	60	90	79	82	161
75th	90	90	90	80	90	90	100	90	70	100	89	87	176
Max	100	100	100	100	100	100	100	100	90	100	99	94	193
Average	83.3	76.1	78.9	69.2	80.6	81.6	86.7	78.1	58.7	86.7	78.0	78.8	156.8
	Medical students (MINF 705B/709 A), *n* = 24												
Min	70	70	50	50	70	50	70	60	40	60	59	NA	NA
25th	80	80	70	60	80	70	80	70	50	90	73	NA	NA
Median	80	80	70	70	90	80	90	70	60	100	79	NA	NA
75th	95	90	80	80	100	85	90	80	65	100	86.5	NA	NA
Max	100	100	100	90	100	90	100	100	80	100	96	NA	NA
Average	87.2	80.2	78.2	71.5	87.4	79.0	85.3	80.7	58.6	92.8	80.1	NA	NA
	All students, *n* = 139 for MCQs, *n* = 51 for final exam												
Min	40	40	30	20	20	20	40	20	20	50	30	43	73
25th	80	70	70	60	80	80	80	70	50	90	73	76	149
Median	90	80	80	70	90	80	90	80	60	90	81	85	166
75th	90	90	90	80	100	90	100	90	70	100	90	89.5	179.5
Max	100	100	100	100	100	100	100	100	100	100	100	97	197
Average	85.5	77.7	79.1	70.6	84.8	81.8	86.7	81.0	60.6	89.7	79.7	81.7	161.4

Table 3 | Scores on each unit assessment and the final exam for all LLMs assessed

Student ID	Unit 1	Unit 2	Unit 3	Unit 4	Unit 5	Unit 6	Unit 7	Unit 8	Unit 9	Unit 10	MCQ Average	Final Exam	MCQ+Final Combined
ChatGPT Plus	100	100	80	80	90	100	80	100	70	80	88	76	164
Claude 3 Opus	100	80	70	100	80	100	70	100	40	70	81	91	172
CoPilot Bing-Precise	100	90	80	100	90	100	100	100	70	50	88	85	173
Gemini Pro	100	90	70	90	90	100	90	80	60	80	85	91	176
Llama 3.1 405B	100	100	70	100	100	90	70	100	60	60	85	88	173
Mistral-Large	100	90	80	90	90	80	80	80	60	80	83	82	165

Precise close behind. Gemini Pro scored at about the 75th percentile of all students who had taken the class between early 2023 and early 2024. Although the graduate and continuing education offerings of the courses have additional requirements for their complete grade, the performance of all of the LLMs was well above the passing levels for the MCQ and final exam components of the course. The clock time for the LLMs varied—mainly due to the amount of text printed to the browser window—but was far less than the time typically taken by students, e.g., up to four hours allowed for the latter to complete the final exam. An observation made when grading LLM final exams is that the LLM followed instructions for at most 2 sentence answers and rarely input one-word answers. In contrast, students usually vary the length of their answers and often give one-word answers. Another minor difference is that LLMs complete grammatically correct sentences and have correct spelling all of the time compared to some of the students not responding that way.

The results of this study raise significant questions for the future of student assessment in most if not all academic disciplines. Clearly LLMs can generate output at a high level for graduate-level courses such as introductory biomedical and health informatics. What are the options for maintaining the ability to assess students? One challenge for a course like this is that its focus and assessments are knowledge-based. The course does not develop or assess skills, but instead provides the knowledge and vocabulary for further skills development. This course might also consider, at least for the final examination, abandoning its open-book format.

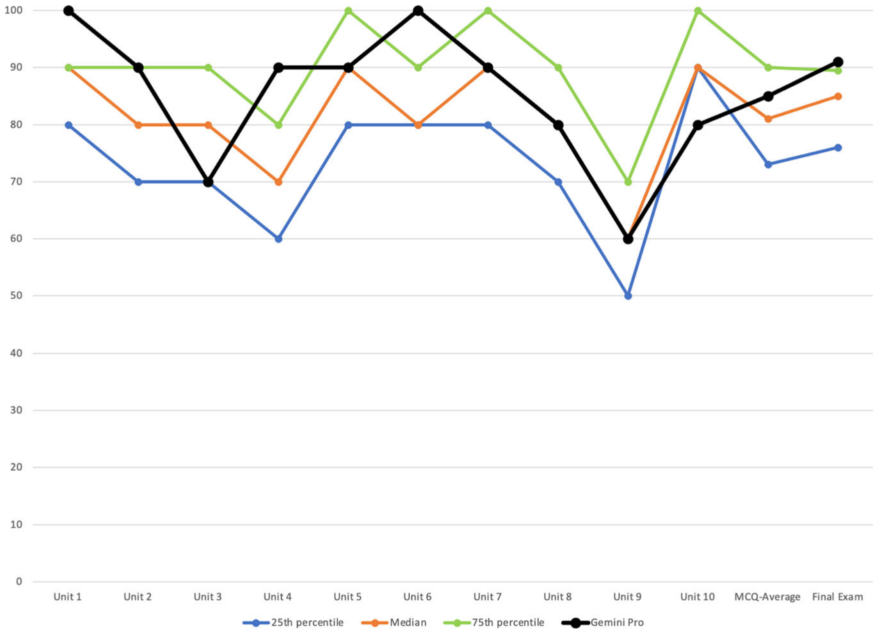


Fig. 1 | Unit assessments and final exam results for students and Gemini Pro. Summary of unit assessments and final exam results for all students at the 25th, 50th (median), and 75th quartile of performance (thinner green, orange, and blue lines respectively) with best-performing LLM, Gemini Pro (thicker black line).

Table 4 | Clock time taken in seconds for each LLM for each assessment

Assessment	ChatGPT Plus (GPT-4)	Claude 3 Opus	CoPilot with Bing-Precise	Gemini Pro	Llama 3.1 405B	Mistral-Large
Unit 1	39	25	24	31	14	15
Unit 2	51	41	25	23	23	16
Unit 3	39	31	14	24	37	11
Unit 4	70	38	19	21	18	12
Unit 5	57	34	18	19	14	11
Unit 6	61	36	22	21	15	13
Unit 7	77	35	20	22	13	15
Unit 8	55	36	23	25	23	14
Unit 9	100	44	26	21	15	17
Unit 10	43	43	25	21	16	29
Final Exam	73	49	80	25	20	38
Total	665	412	296	253	208	191

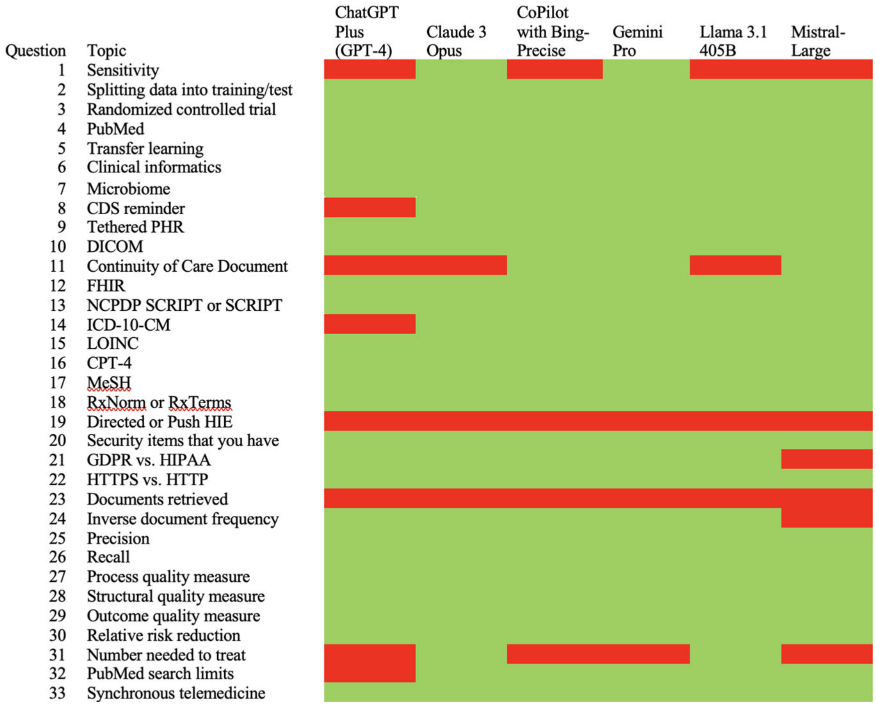
Other options for maintaining the ability to assess students might be to develop more complex "Google-proof" questions for the assessments[32]. Some suggest the use of generative AI detectors, although a review of recent research found mixed ability to detect text coming from LLMs[33]. One concern for such detectors is their propensity to misclassify non-native English writing as generated by AI[34].

The success of LLMs in educational tasks has implications beyond the student phase of education. If students are able to excel in classes due to generative AI, this may impact professional practice of graduates who have not necessarily mastered the foundational knowledge of fields in which they work. Assessments may be particularly problematic for adult learners who take mostly online courses asynchronously and cannot come to campuses for proctored exams. Indeed, Cooper and Rodman note that LLM use in medical education has "the potential to be at least as disruptive as the problem-oriented medical record, having passed both licensing and clinical reasoning exams and approximating the diagnostic thought patterns of physicians."[35] Mollick notes that educators face a "homework apocalypse" in simple prompting of LLMs being able to achieve passing or even better grades on assessments[36].

There were a number of limitations to this study. First, we reviewed LLM performance in a single course and the results may not generalize to other graduate, continuing education, and/or medical student courses. Second, students after the November 2022 release of ChatGPT may have used generative AI themselves in the course, which could have had beneficial or detrimental effect on their performance. Third, since the biomedical and health informatics field evolves rapidly, including in but not limited to AI, how performance in courses on it is impacted in the long run by LLMs is unknown. Finally, there are reproducibility challenges for using industry-provided LLMs, although this is true for just about all studies using such LLMs, which undergo constant change and updating. We do not, however, believe that these limitations undermine our main results and conclusion, which is that LLMs scored at between the 50th and 75th percentile for a highly subscribed introductory biomedical informatics course.

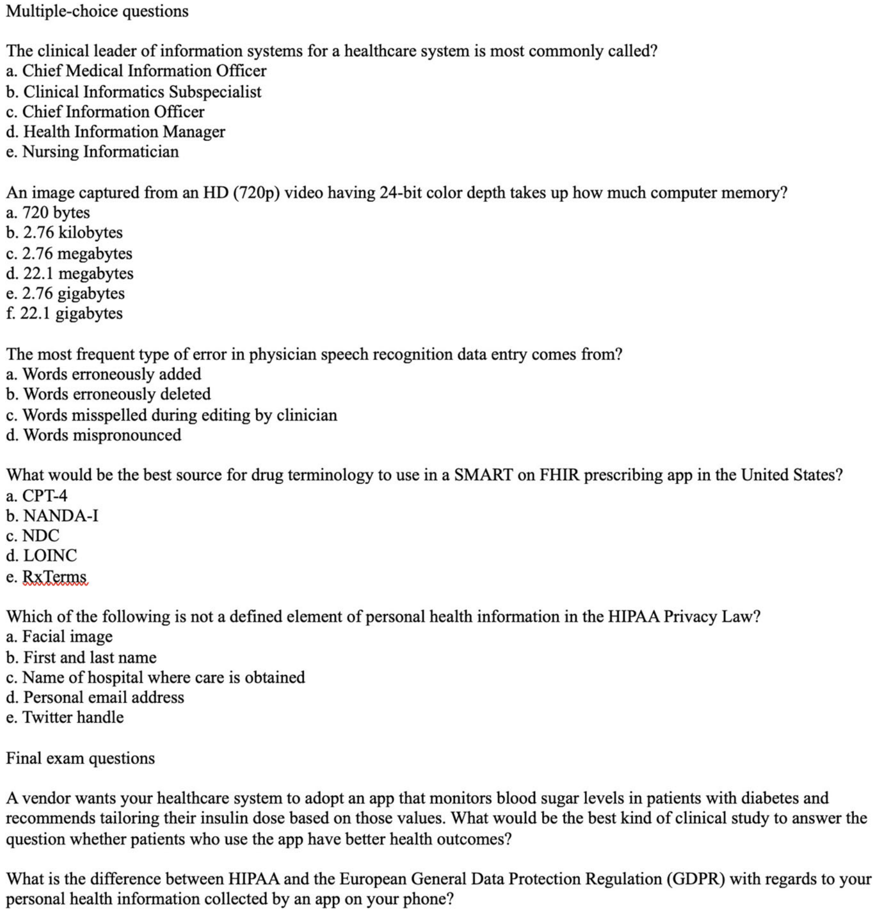
In conclusion, we found that the best LLM system exceeded the performance of about three-quarters of graduate, continuing education, and medical students taking an introductory online course in biomedical and health informatics. Our results showed that LLMs are

Fig. 2 | Correct and incorrect answers on final exam for all LLMs. Topics with correct (green) and incorrect (red) answers on final exam for all LLMs.



Question	Topic	ChatGPT Plus (GPT-4)	Claude 3 Opus	CoPilot with Bing-Precise	Gemini Pro	Llama 3.1 405B	Mistral-Large
1	Sensitivity	🟥	🟩	🟥	🟩	🟥	🟥
2	Splitting data into training/test	🟩	🟩	🟩	🟩	🟩	🟩
3	Randomized controlled trial	🟩	🟩	🟩	🟩	🟩	🟩
4	PubMed	🟩	🟩	🟩	🟩	🟩	🟩
5	Transfer learning	🟩	🟩	🟩	🟩	🟩	🟩
6	Clinical informatics	🟩	🟩	🟩	🟩	🟩	🟩
7	Microbiome	🟩	🟩	🟩	🟩	🟩	🟩
8	CDS reminder	🟥	🟩	🟩	🟩	🟩	🟩
9	Tethered PHR	🟩	🟩	🟩	🟩	🟩	🟩
10	DICOM	🟩	🟩	🟩	🟩	🟩	🟩
11	Continuity of Care Document	🟥	🟥	🟩	🟩	🟥	🟩
12	FHIR	🟩	🟩	🟩	🟩	🟩	🟩
13	NCPDP SCRIPT or SCRIPT	🟩	🟩	🟩	🟩	🟩	🟩
14	ICD-10-CM	🟥	🟩	🟩	🟩	🟩	🟩
15	LOINC	🟩	🟩	🟩	🟩	🟩	🟩
16	CPT-4	🟩	🟩	🟩	🟩	🟩	🟩
17	MeSH	🟩	🟩	🟩	🟩	🟩	🟩
18	RxNorm or RxTerms	🟩	🟩	🟩	🟩	🟩	🟩
19	Directed or Push HIE	🟥	🟥	🟥	🟥	🟥	🟥
20	Security items that you have	🟩	🟩	🟩	🟩	🟩	🟩
21	GDPR vs. HIPAA	🟩	🟩	🟩	🟩	🟩	🟥
22	HTTPS vs. HTTP	🟩	🟩	🟩	🟩	🟩	🟩
23	Documents retrieved	🟥	🟥	🟥	🟥	🟥	🟥
24	Inverse document frequency	🟩	🟩	🟩	🟩	🟩	🟥
25	Precision	🟩	🟩	🟩	🟩	🟩	🟩
26	Recall	🟩	🟩	🟩	🟩	🟩	🟩
27	Process quality measure	🟩	🟩	🟩	🟩	🟩	🟩
28	Structural quality measure	🟩	🟩	🟩	🟩	🟩	🟩
29	Outcome quality measure	🟩	🟩	🟩	🟩	🟩	🟩
30	Relative risk reduction	🟩	🟩	🟩	🟩	🟩	🟩
31	Number needed to treat	🟥	🟩	🟥	🟥	🟩	🟥
32	PubMed search limits	🟥	🟩	🟩	🟩	🟩	🟩
33	Synchronous telemedicine	🟩	🟩	🟩	🟩	🟩	🟩

Fig. 3 | Multiple-choice and final exams questions. Example multiple-choice and final exams questions used in this study.



Multiple-choice questions

The clinical leader of information systems for a healthcare system is most commonly called?
a. Chief Medical Information Officer
b. Clinical Informatics Subspecialist
c. Chief Information Officer
d. Health Information Manager
e. Nursing Informatician

An image captured from an HD (720p) video having 24-bit color depth takes up how much computer memory?
a. 720 bytes
b. 2.76 kilobytes
c. 2.76 megabytes
d. 22.1 megabytes
e. 2.76 gigabytes
f. 22.1 gigabytes

The most frequent type of error in physician speech recognition data entry comes from?
a. Words erroneously added
b. Words erroneously deleted
c. Words misspelled during editing by clinician
d. Words mispronounced

What would be the best source for drug terminology to use in a SMART on FHIR prescribing app in the United States?
a. CPT-4
b. NANDA-I
c. NDC
d. LOINC
e. RxTerms

Which of the following is not a defined element of personal health information in the HIPAA Privacy Law?
a. Facial image
b. First and last name
c. Name of hospital where care is obtained
d. Personal email address
e. Twitter handle

Final exam questions

A vendor wants your healthcare system to adopt an app that monitors blood sugar levels in patients with diabetes and recommends tailoring their insulin dose based on those values. What would be the best kind of clinical study to answer the question whether patients who use the app have better health outcomes?

What is the difference between HIPAA and the European General Data Protection Regulation (GDPR) with regards to your personal health information collected by an app on your phone?

having a profound effect on education and its assessment. Certainly, LLMs will be part of the toolkit of professionals and academics in all disciplines. The challenge is how LLM use from the beginning of learning may impact mastery of competence and professional behavior later on. Future research must address these concerns to determine the optimal role of generative AI in all levels of education.

Methods

We compared student performance on MCQs and the final exam with 6 state-of-the-art LLMs: ChatGPT Plus (GPT-4), Claude 3 Opus, CoPilot with Bing-Precise, Gemini Pro, Llama 3.1 405B, and Mistral-Large. Use of de-identified aggregate student scores was determined by the OHSU Institutional Review Board (IRB) to be research not involving human subjects, with IRB review and approval not required (STUDY00026901). This enabled us to calculate the average grades for students in the different offerings of the course using the 2023 content. We used Microsoft Excel to calculate median and related scores.

To assess LLM performance, we used the latest versions of LLMs in their user-interactive modes since this was likely how most students would access them. Each LLM was prompted by a standard approach:

- MCQs: Each LLM was prompted first with, “You are a graduate student taking an introductory course in biomedical and health informatics. Please provide the best answers to the following multiple-choice questions.” This was followed by pasting in the MCQs one unit (10 questions) at a time exactly as they appeared in the MCQ preview file in the Sakai LMS.
- Final exam: Each LLM was prompted with, “You are a graduate student taking the final exam in an introductory course in biomedical and health informatics. Answer each of the following questions with a short answer that is one sentence or less.” This was followed by pasting in the exam, which had 33 questions, separated into 8 sections with a one-sentence heading for each section, exactly as it appeared in the Sakai LMS exam module.

The LLM models used were prompted on the following days and times using their standard interactive interfaces:

- ChatGPT Plus (GPT-4) on February 20, 2024 at 3 pm Pacific Standard Time (PST)
- Gemini Pro on February 28, 2024 at 4 pm PST
- Mistral-Large on March 1, 2024 at 3 pm PST
- CoPilot with Bing-Precise on March 1, 2024 at 4 pm PST
- Claude 3 Opus on March 10, 2024 at 6 am Pacific Daylight Time (PDT)
- Llama 3.1 405B on August 16, 2024 at 1 pm PDT

We captured the text output from each of the LLMs, and these were manually graded by KFH and reviewed by WRH. The prompts and answer keys are provided in Supplementary Notes 2-5. Some sample questions are shown in Fig. 3.

The analysis had two small amounts of missing data unlikely to impact the overall results. Data from two of the 10 units for BMI 510/610 and AMIA 10 × 10 were not used for the last group of students (6 in BMI 510/610 and 43 in AMIA 10 × 10) taking the course in late 2023-early 2024 because some of the course content was updated requiring updating of the MCQs for those units. In addition, a configuration error in the Sakai LMS lost the individual but not aggregate quiz results for 6 students taking 705B/709 A in early 2024.

Data availability

All data generated or analyzed during this study are included in this published article and its supplementary information file.

Received: 18 April 2024; Accepted: 3 September 2024;
Published online: 13 September 2024

References

1. Omiye, J. A., Gui, H., Rezaei, S. J., Zou, J. & Daneshjou, R. Large Language Models in Medicine: The Potentials and Pitfalls: A Narrative Review. *Ann. Intern Med*. **177**, 210–220 (2024).
2. Bhayana, R., Bleakney, R. R. & Krishna, S. GPT-4 in Radiology: Improvements in Advanced Reasoning. *Radiology* **307**, e230987 (2023).
3. Kumah-Crystal, Y., Mankowitz, S., Embi, P. & Lehmann, C. U. ChatGPT and the clinical informatics board examination: the end of unproctored maintenance of certification? *J Am Med Inform Assoc* ocad104 (2023) https://doi.org/10.1093/jamia/ocad104.
4. Nori, H. et al. Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine. Preprint at https://doi.org/10.48550/arXiv.2311.16452 (2023).
5. Katz, U. et al. GPT versus Resident Physicians — A Benchmark Based on Official Board Scores. *NEJM AI* **0**, AIdbp2300192 (2024).
6. Li, S. W. et al. ChatGPT outscored human candidates in a virtual objective structured clinical examination in obstetrics and gynecology. *Am J Obstet Gynecol* S0002-9378(23)00251-X (2023) https://doi.org/10.1016/j.ajog.2023.04.020.
7. Dash, D. et al. Evaluation of GPT-3.5 and GPT-4 for supporting real-world information needs in healthcare delivery. Preprint at https://doi.org/10.48550/arXiv.2304.13714 (2023).
8. Goodman, R. S. et al. Accuracy and Reliability of Chatbot Responses to Physician Questions. *JAMA Netw. Open* **6**, e2336483 (2023).
9. Benoit, J. R. A. ChatGPT for Clinical Vignette Generation, Revision, and Evaluation. 2023.02.04.23285478 Preprint at https://doi.org/10.1101/2023.02.04.23285478 (2023).
10. Eriksen, A. V., Möller, S. & Ryg, J. Use of GPT-4 to Diagnose Complex Clinical Cases. *NEJM AI* (2023) https://doi.org/10.1056/AIp2300031.
11. Kanjee, Z., Crowe, B. & Rodman, A. Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge. *JAMA* **330**, 78–80 (2023).
12. Levine, D. M. et al. The Diagnostic and Triage Accuracy of the GPT-3 Artificial Intelligence Model: An Observational Study. *Lancet Digit Health.* **6**, e555–e561 (2024).
13. McDuff, D. et al. Towards Accurate Differential Diagnosis with Large Language Models. Preprint at https://doi.org/10.48550/arXiv.2312.00164 (2023).
14. Tu, T. et al. Towards Conversational Diagnostic AI. Preprint at https://doi.org/10.48550/arXiv.2401.05654 (2024).
15. Cabral, S. et al. Clinical Reasoning of a Generative Artificial Intelligence Model Compared With Physicians. *JAMA Intern Med* https://doi.org/10.1001/jamainternmed.2024.0295 (2024).
16. Mollick, E. R. & Mollick, L. Using AI to Implement Effective Teaching Strategies in Classrooms: Five Strategies, Including Prompts. SSRN Scholarly Paper at https://doi.org/10.2139/ssrn.4391243 (2023).
17. Denny, P. et al. Computing Education in the Era of Generative AI. *Commun. ACM* **67**, 56–67 (2024).
18. Johnson, M. Generative AI and CS Education. *Commun. ACM* **67**, 23–24 (2024).
19. Poldrack, R. A., Lu, T. & Beguš, G. AI-assisted coding: Experiments with GPT-4. Preprint at https://doi.org/10.48550/arXiv.2304.13187 (2023).
20. Choi, J. H., Monahan, A. & Schwarcz, D. Lawyering in the Age of Artificial Intelligence. SSRN Scholarly Paper at https://doi.org/10.2139/ssrn.4626276 (2023).
21. Hong, S. et al. Data Interpreter: An LLM Agent For Data Science. Preprint at https://doi.org/10.48550/arXiv.2402.18679 (2024).
22. Hersh, W. Search still matters: information retrieval in the era of generative AI. *J Am Med Inform Assoc* ocae014 (2024) https://doi.org/10.1093/jamia/ocae014.

23. Wu, K. et al. How well do LLMs cite relevant medical references? An evaluation framework and analyses. Preprint at https://doi.org/10.48550/arXiv.2402.02008 (2024).
24. Augenstein, I. et al. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nat Mach Intell.* **6**, 852–863 (2024).
25. Heston, T. F. & Khun, C. Prompt Engineering in Medical Education. *Int. Med. Educ.* **2**, 198–205 (2023).
26. Mollick, E. R. & Mollick, L. Assigning AI: Seven Approaches for Students, with Prompts. SSRN Scholarly Paper at https://doi.org/10.2139/ssrn.4475995 (2023).
27. Preiksaitis, C. & Rose, C. Opportunities, Challenges, and Future Directions of Generative Artificial Intelligence in Medical Education: Scoping Review. *JMIR Med Educ.* **9**, e48785 (2023).
28. Sok, S. & Heng, K. ChatGPT for Education and Research: A Review of Benefits and Risks. SSRN Scholarly Paper at https://doi.org/10.2139/ssrn.4378735 (2023).
29. Huschens, M., Briesch, M., Sobania, D. & Rothlauf, F. Do You Trust ChatGPT? -- Perceived Credibility of Human and AI-Generated Content. Preprint at https://doi.org/10.48550/arXiv.2309.02524 (2023).
30. Hersh, W. Competencies and Curricula Across the Spectrum of Learners for Biomedical and Health Informatics. *Stud. Health Technol. Inf.* **300**, 93–107 (2022).
31. Hersh, W. & Williamson, J. Educating 10,000 informaticians by 2010: the AMIA 10x10 program. *Int J. Med Inf.* **76**, 377–382 (2007).
32. Rein, D. et al. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. Preprint at https://doi.org/10.48550/arXiv.2311.12022 (2023).
33. Tang, R., Chuang, Y.-N. & Hu, X. The Science of Detecting LLM-Generated Text. *Commun. ACM* **67**, 50–59 (2024).
34. Liang, W., Yuksekgonul, M., Mao, Y., Wu, E. & Zou, J. GPT detectors are biased against non-native English writers. *Patterns (N. Y)* **4**, 100779 (2023).
35. Cooper, A. & Rodman, A. AI and Medical Education — A 21st-Century Pandora's Box. *New England Journal of Medicine* (2023) https://doi.org/10.1056/NEJMp2304993.
36. Mollick, E. The Homework Apocalypse. *One Useful Thing* https://www.oneusefulthing.org/p/the-homework-apocalypse (2023).

Author contributions

W.R.H. carried out LLM prompting, calculated aggregate student scores, and carried out comparisons with LLM scores. He wrote the manuscript, with review by K.F.H. K.F.H. graded MCQ and final exam answers by the LLMs, with review by W.R.H.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-024-01251-0.

Correspondence and requests for materials should be addressed to William Hersh.

Reprints and permissions information is available at http://www.nature.com/reprints