Accomplishments and Limitations of LLMs in Education

Medinfo 2025 Panel Presentation
Taipei, Taiwan
August 10, 2025

William Hersh, MD
Professor
Division of Informatics, Clinical Epidemiology, and Translational Data Science
Department of Medicine
School of Medicine
Oregon Health & Science University
Portland, OR, USA
Email: hersh@ohsu.edu
Web: https://billhersh.info/
BlueSky: @billhersh.bsky.social

References

Al-Sibai, N., 2025. College Students Are Sprinkling Typos Into Their AI Papers on Purpose [WWW Document]. Futurism. URL https://futurism.com/college-students-ai-typos (accessed 7.21.25).

Avramovic, S., Avramovic, I., Wojtusiak, J., 2024. Exploring the Impact of GitHub Copilot on Health Informatics Education. Appl Clin Inform 15, 1121–1129. https://doi.org/10.1055/a-2414-7790

Gallant, T.B., Rettinger, D.A., 2025. The Opposite of Cheating: Teaching for Integrity in the Age of AI. University of Oklahoma Press.

Hersh, W., 2025. Generative Artificial Intelligence: Implications for Biomedical and Health Professions Education. Annu Rev Biomed Data Sci. https://doi.org/10.1146/annurev-biodatasci-103123-094756

Hersh, W., 2024a. A Quarter-Century of Online Informatics Education: Learners Served and Lessons Learned. J Med Internet Res 26, e59066. https://doi.org/10.2196/59066

Hersh, W., 2024b. Search still matters: information retrieval in the era of generative AI. J Am Med Inform Assoc 31, 2159–2161. https://doi.org/10.1093/jamia/ocae014

Hersh, W., 2022. 3000 by 2022 - A New Milestone for the 10x10 Course. Informatics Professor. URL https://informaticsprofessor.blogspot.com/2022/04/3000-by-2022-new-milestone-for-10x10.html (accessed 5.19.22).

Hersh, W., Fultz Hollis, K., 2024. Results and implications for generative AI in a large introductory biomedical and health informatics course. NPJ Digit Med 7, 247. https://doi.org/10.1038/s41746-024-01251-0

Hersh, W., Williamson, J., 2007. Educating 10,000 informaticians by 2010: the AMIA 10x10 program. Int J Med Inform 76, 377–382. https://doi.org/10.1016/j.ijmedinf.2007.01.003

Kosmyna, N., Hauptmann, E., Yuan, Y.T., Situ, J., Liao, X.-H., Beresnitzky, A.V., Braunstein, I., Maes, P., 2025. Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task. https://doi.org/10.48550/arXiv.2506.08872

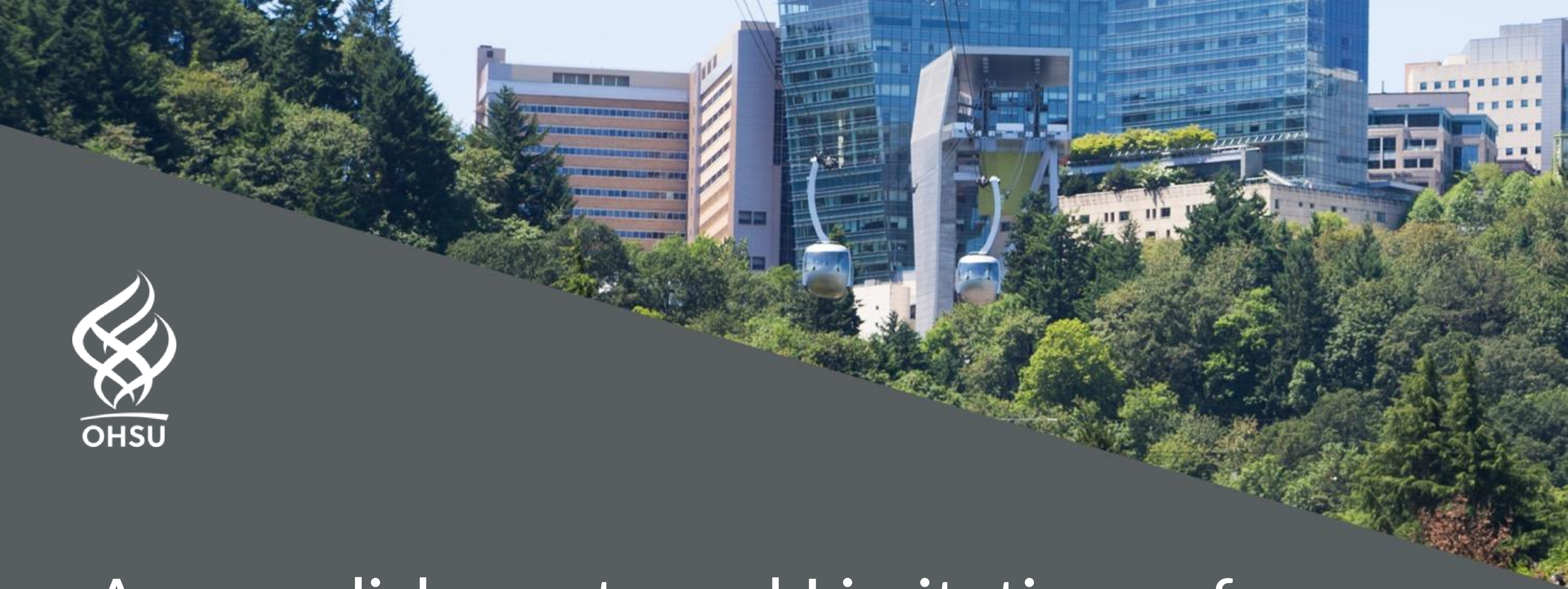
Lee, H.-P. (Hank), Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., Wilson, N., 2025. The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers, in: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI '25. Association for Computing Machinery, New York, NY, USA, pp. 1–22. https://doi.org/10.1145/3706598.3713778

McMurtrie, B., 2024. Cheating Has Become Normal [WWW Document]. The Chronicle of Higher Education. URL https://www.chronicle.com/article/cheating-has-become-normal (accessed 7.21.25).

Mollick, E., 2023. The Homework Apocalypse [WWW Document]. One Useful Thing. URL https://www.oneusefulthing.org/p/the-homework-apocalypse (accessed 3.20.24).

Norlen, N., Barrett, G., 2023. Word of the Year 2023 [WWW Document]. Dictionary.com. URL https://content.dictionary.com/word-of-the-year-2023/ (accessed 10.9.24).

Stribling, D., Xia, Y., Amer, M.K., Graim, K.S., Mulligan, C.J., Renne, R., 2024. The model student: GPT-4 performance on graduate biomedical science exams. Sci Rep 14, 5670. https://doi.org/10.1038/s41598-024-55568-7



Accomplishments and Limitations of LLMs in Education

William Hersh
Professor
Oregon Health & Science University
Portland, OR, USA

Outline

- Accomplishments of LLMs in biomedicine and in education outside biomedicine
- Results of LLMs in biomedical sciences and informatics courses
- Implications and challenges for generative AI in student learning and assessment

Accomplishments of LLMs in biomedicine and education outside biomedicine (Hersh, 2025)

- LLMs perform well in many different types of biomedical tasks
  - Medical board exams
  - Graduate school bioscience exams
  - Objective structured clinical exams (OSCEs)
  - Answering clinical questions
  - Solving clinical cases
  - Conversational diagnostic dialogue
  - Clinical reasoning

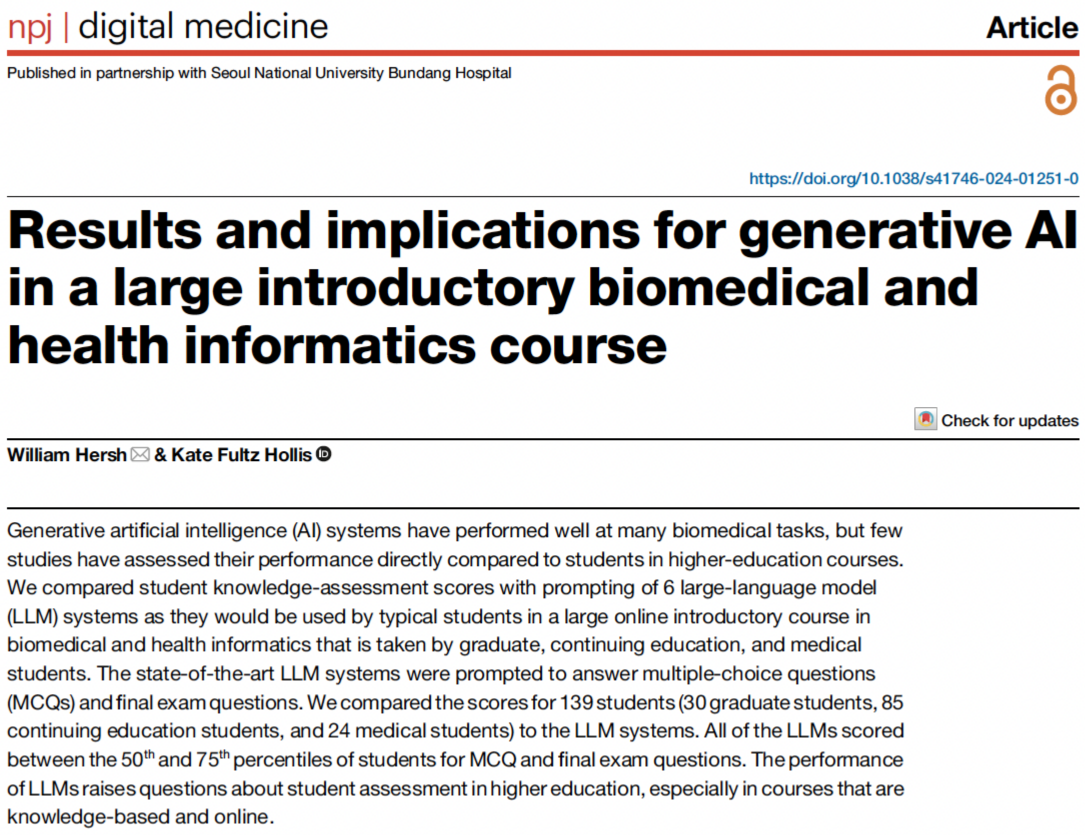
- LLMs perform well in education in disciplines outside biomedicine
  - High school standardized and AP exams
  - Computer programming
  - Data science pipelines
  - Predicting company earnings
  - Legal briefs
  - PhD-level biology, chemistry, and physics

Results of LLMs in biomedical sciences and informatics courses

- Introductory biomedical informatics course
- Biosciences exams
- Health informatics programming

Generative AI in an introductory informatics course (Hersh and Fultz Hollis, 2024)

- Large introductory biomedical and health informatics course
  - Same curriculum and (mostly) assessments in courses taught to graduate students, medical students, and continuing education students
- Generative AI
  - Use of large language models (LLMs) in knowledge assessment
- Results and implications
  - How do LLMs fare on student assessments?
  - What does this mean for student assessment in this and other similar courses?



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.

About the course

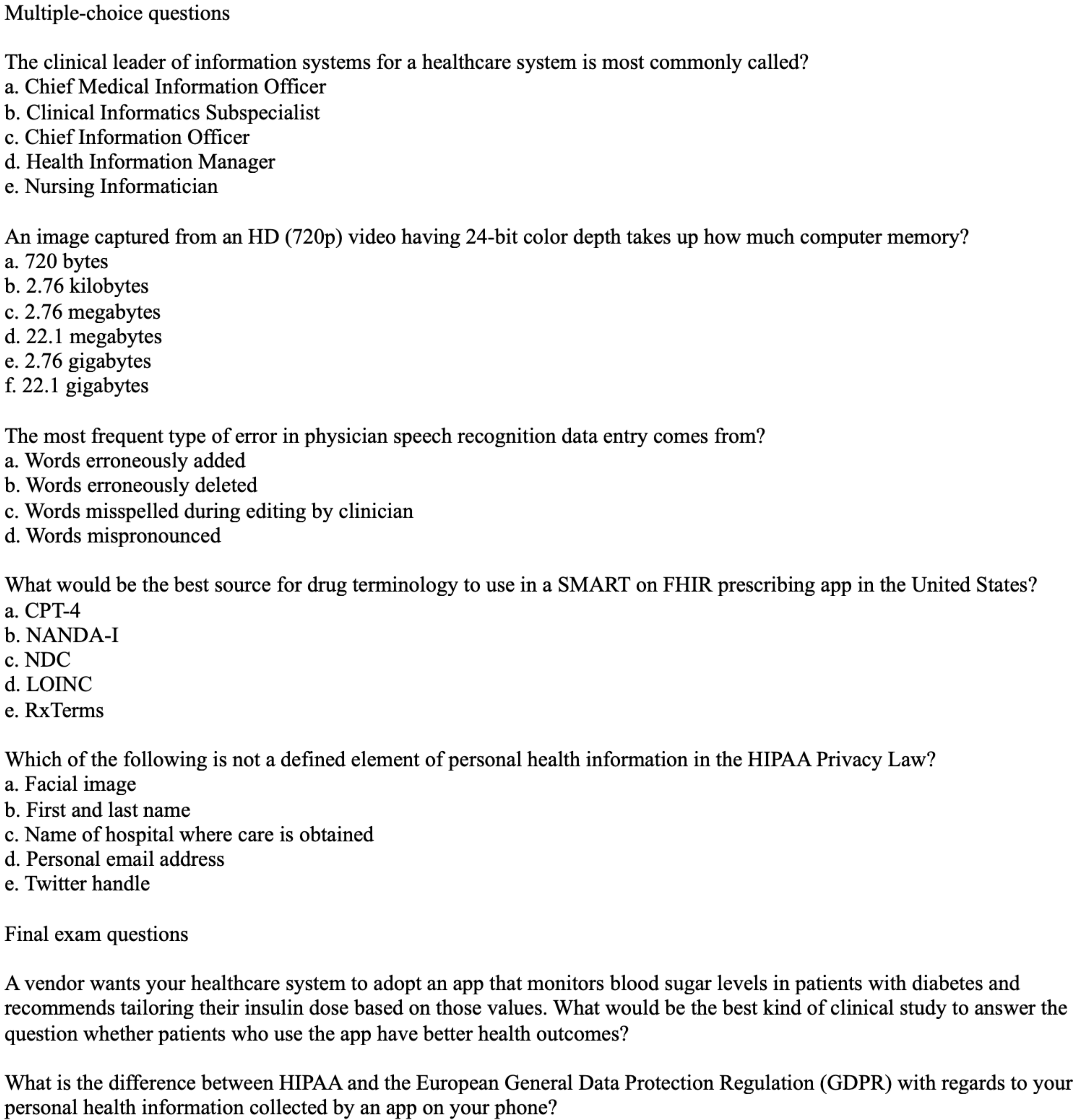
- Introductory overview of biomedical and health informatics (Hersh, 2007; Hersh, 2022)
  - Taught at OHSU for over three decades (Hersh, 2024)
  - Updated annually
  - Also known as 10x10 (“ten by ten”) for continuing education students
- Taught online using
  - Voice-over-Powerpoint lectures
  - Discussion forums
  - Optional textbook readings
- Assessments include
  - Multiple-choice questions (MCQs) – 10 questions in each of 10 units
  - Final exam – 33 short-answer questions
  - Term paper – not required of medical students, not assessed in this study

Compared 2023 student performance with six commercial, readily available LLMs

- LLMs prompted in Feb-March 2024
  - ChatGPT-4
  - Microsoft CoPilot/Bing – uses GPT-4
  - Google Gemini Pro 1.0
  - Claude 3 Opus
  - Mistral-Large
- Prompted in August 2024
  - Meta Llama 3.1 405B – “open-source”

- Prompted via Web interfaces as students would likely do
- Deemed “non-human research” by IRB

Example questions



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

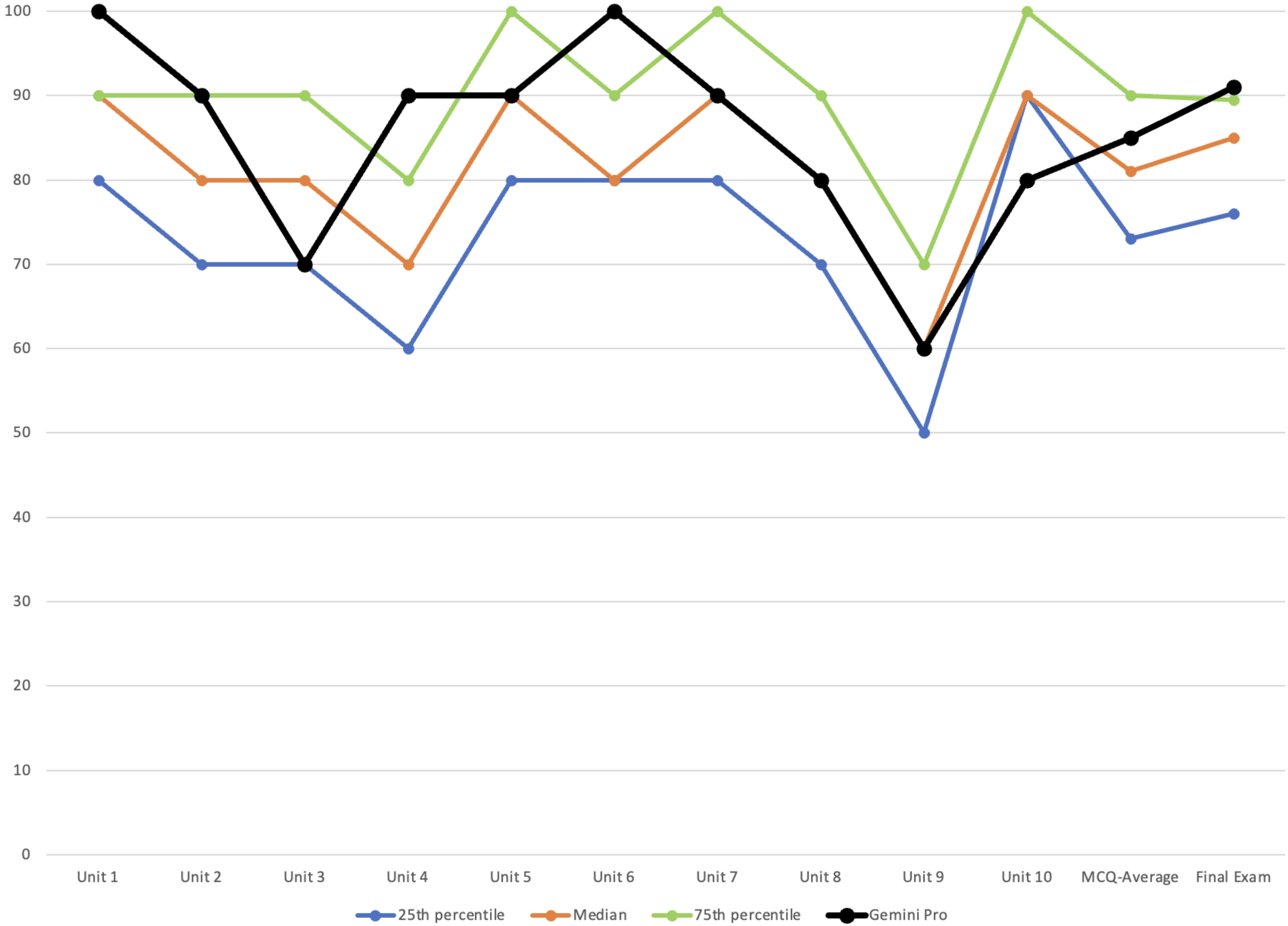
Which of the following is not a defined element of personal health information in the HIPAA Privacy Law?
a. Facial image
b. First and last name
c. Name of hospital where care is obtained
d. Personal email address
e. Twitter handle

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?

Results – overview

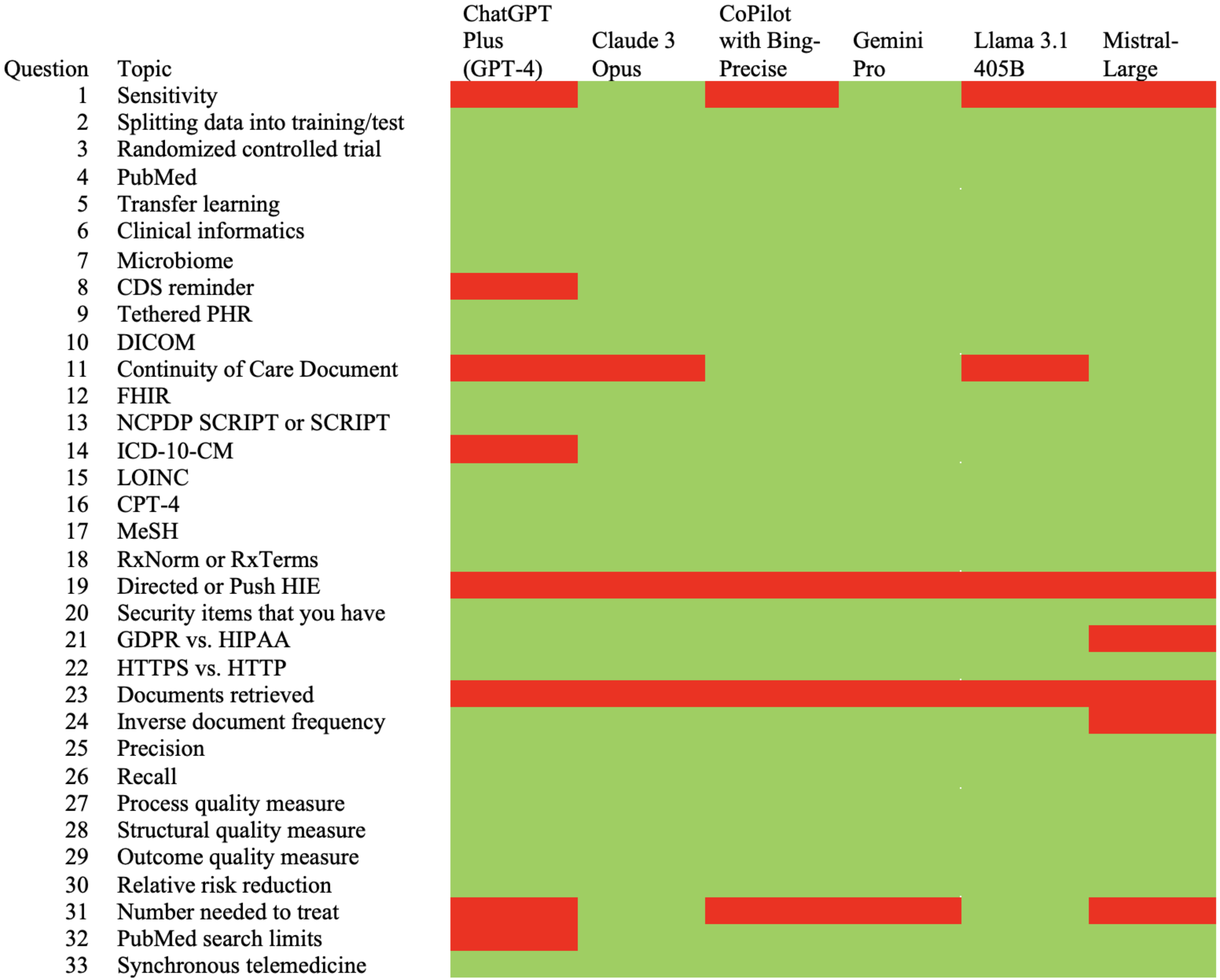


Results – clock time (in seconds)

LLM	MCQ Average	Final Exam
ChatGPT Plus	59.2	73
Claude 3 Opus	36.3	49
CoPilot Bing-Precise	21.6	80
Gemini Pro	22.8	25
Llama 3.1 405B	18.8	20
Mistral-Large	15.3	38

- Most time taken due to screen writing, so actual processing even faster

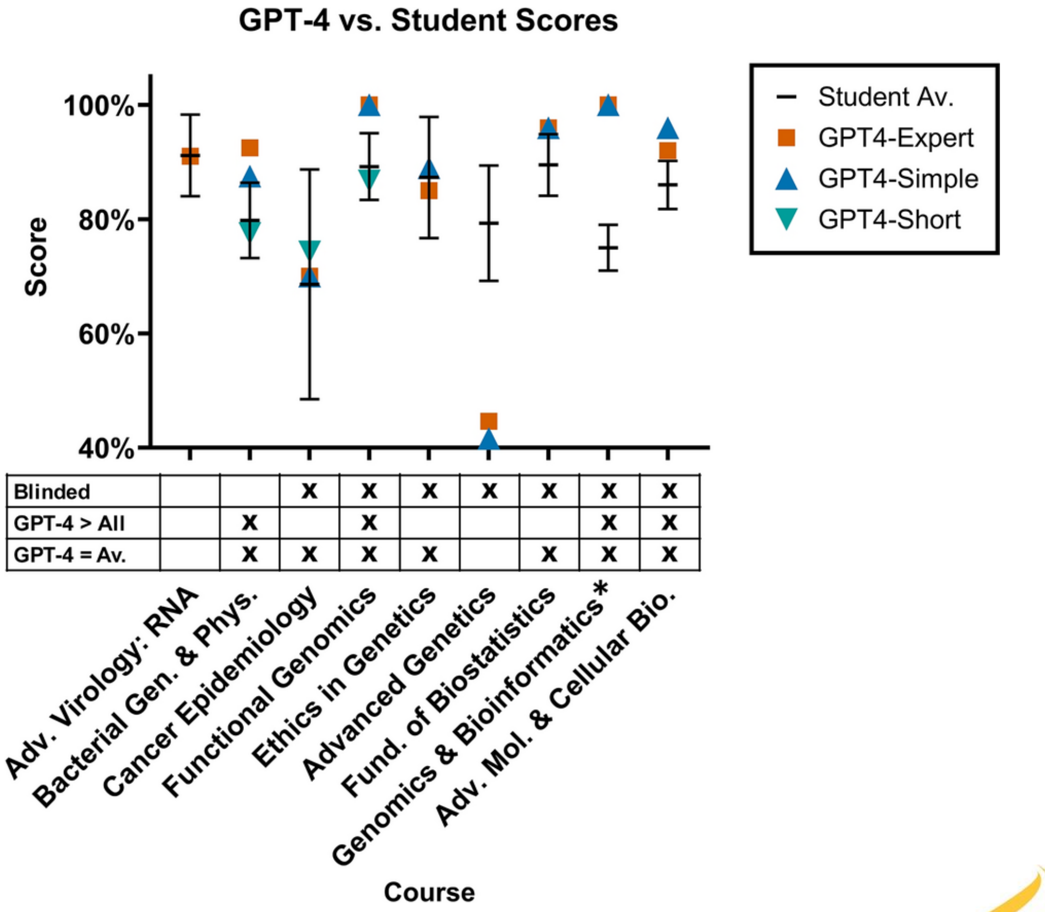
LLMs on final exam questions



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	🟩	🟩	🟩	🟩	🟩	🟩

Related research – graduate-level examinations in biomedical sciences (Stribling, 2024)

- GPT-4 performance on 9 exams
- Exceeded student average on 7 of 9 exams and all student scores for 4 exams
- Performed very well on
  - Fill-in-the-blank, short-answer, and essay questions
  - Questions on figures sourced from published manuscripts
- Performed poorly on questions with
  - With figures containing simulated data
  - Requiring hand-drawn answer

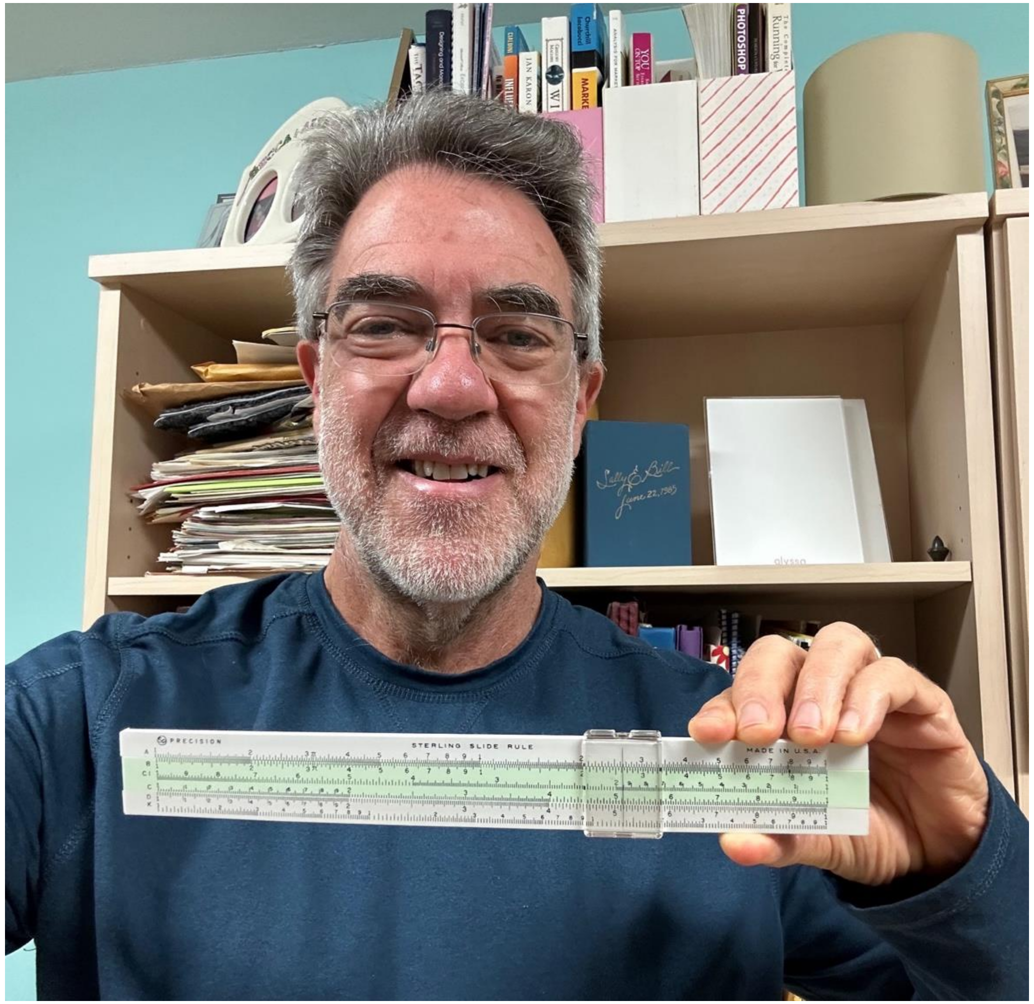


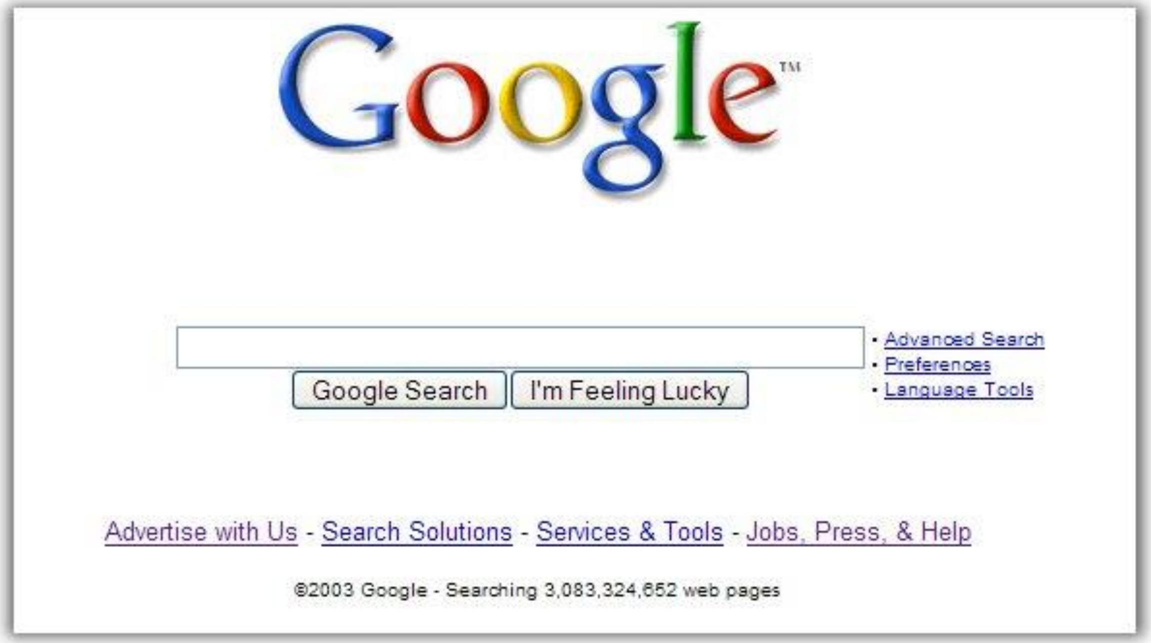
	Adv. Virology: RNA	Bacterial Gen. & Phys.	Cancer Epidemiology	Functional Genomics	Ethics in Genetics	Advanced Genetics	Fund. of Biostatistics	Genomics & Bioinformatics*	Adv. Mol. & Cellular Bio.
Blinded			X	X	X	X	X	X	X
GPT-4 > All		X		X				X	X
GPT-4 = Av.		X	X	X	X		X	X	X

Use of GitHub CoPilot in health informatics programming course (Avramovic, 2024)

- Assessed in problems for
  - Database queries with SQL
  - Computational tasks with Python
- Generated solutions worked well for simple tasks but less well for complex ones
  - Some solutions correct but not most efficient approach

Now what? Educational cusps in my lifetime

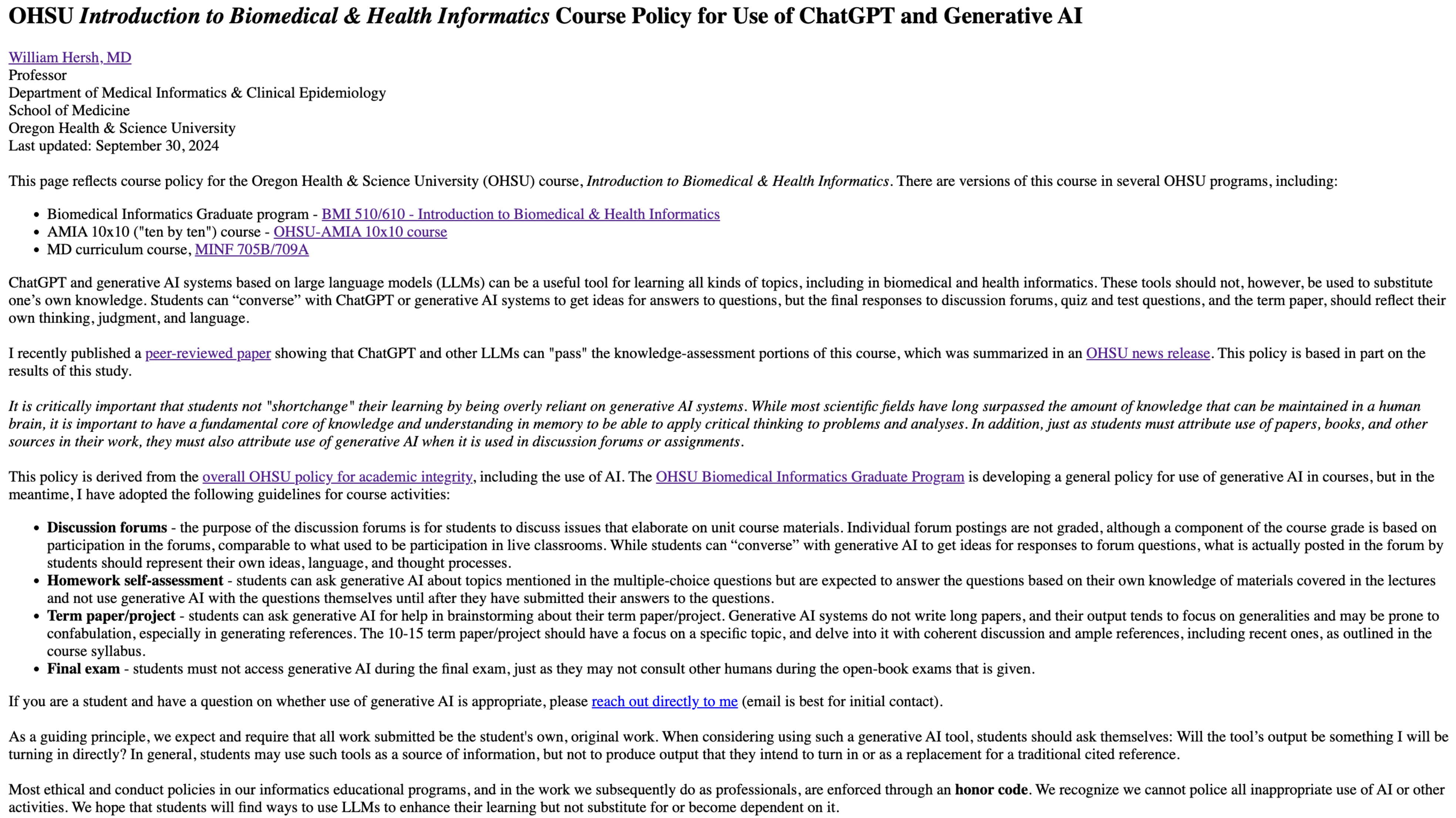




©2003 Google - Searching 3,083,324,652 web pages

First step

https://dmice.ohsu.edu/hersh/introcourse-generativeAI-policy.html



OHSU *Introduction to Biomedical & Health Informatics* Course Policy for Use of ChatGPT and Generative AI

William Hersh, MD
Professor
Department of Medical Informatics & Clinical Epidemiology
School of Medicine
Oregon Health & Science University
Last updated: September 30, 2024

This page reflects course policy for the Oregon Health & Science University (OHSU) course, *Introduction to Biomedical & Health Informatics*. There are versions of this course in several OHSU programs, including:

- Biomedical Informatics Graduate program - BMI 510/610 - Introduction to Biomedical & Health Informatics
- AMIA 10x10 ("ten by ten") course - OHSU-AMIA 10x10 course
- MD curriculum course, MINF 705B/709A

ChatGPT and generative AI systems based on large language models (LLMs) can be a useful tool for learning all kinds of topics, including in biomedical and health informatics. These tools should not, however, be used to substitute one's own knowledge. Students can "converse" with ChatGPT or generative AI systems to get ideas for answers to questions, but the final responses to discussion forums, quiz and test questions, and the term paper, should reflect their own thinking, judgment, and language.

I recently published a peer-reviewed paper showing that ChatGPT and other LLMs can "pass" the knowledge-assessment portions of this course, which was summarized in an OHSU news release. This policy is based in part on the results of this study.

It is critically important that students not "shortchange" their learning by being overly reliant on generative AI systems. While most scientific fields have long surpassed the amount of knowledge that can be maintained in a human brain, it is important to have a fundamental core of knowledge and understanding in memory to be able to apply critical thinking to problems and analyses. In addition, just as students must attribute use of papers, books, and other sources in their work, they must also attribute use of generative AI when it is used in discussion forums or assignments.

This policy is derived from the overall OHSU policy for academic integrity, including the use of AI. The OHSU Biomedical Informatics Graduate Program is developing a general policy for use of generative AI in courses, but in the meantime, I have adopted the following guidelines for course activities:

- **Discussion forums** - the purpose of the discussion forums is for students to discuss issues that elaborate on unit course materials. Individual forum postings are not graded, although a component of the course grade is based on participation in the forums, comparable to what used to be participation in live classrooms. While students can "converse" with generative AI to get ideas for responses to forum questions, what is actually posted in the forum by students should represent their own ideas, language, and thought processes.
- **Homework self-assessment** - students can ask generative AI about topics mentioned in the multiple-choice questions but are expected to answer the questions based on their own knowledge of materials covered in the lectures and not use generative AI with the questions themselves until after they have submitted their answers to the questions.
- **Term paper/project** - students can ask generative AI for help in brainstorming about their term paper/project. Generative AI systems do not write long papers, and their output tends to focus on generalities and may be prone to confabulation, especially in generating references. The 10-15 term paper/project should have a focus on a specific topic, and delve into it with coherent discussion and ample references, including recent ones, as outlined in the course syllabus.
- **Final exam** - students must not access generative AI during the final exam, just as they may not consult other humans during the open-book exams that is given.

If you are a student and have a question on whether use of generative AI is appropriate, please reach out directly to me (email is best for initial contact).

As a guiding principle, we expect and require that all work submitted be the student's own, original work. When considering using such a generative AI tool, students should ask themselves: Will the tool's output be something I will be turning in directly? In general, students may use such tools as a source of information, but not to produce output that they intend to turn in or as a replacement for a traditional cited reference.

Most ethical and conduct policies in our informatics educational programs, and in the work we subsequently do as professionals, are enforced through an **honor code**. We recognize we cannot police all inappropriate use of AI or other activities. We hope that students will find ways to use LLMs to enhance their learning but not substitute for or become dependent on it.

Concerns for LLMs in education

- Prone to hallucinations, confafabulations, etc.
    - – Dictionary.com 2023 word of year: hallucinate (Norlen, 2023)
- Is there a “homework apocalypse” (Mollick, 2023)?
- Provide answers but not sources of knowledge (Hersh, 2024)
- “Cheating has become new normal” (McMurtrie, 2024)
- “If you’re going to cheat, you have to cheat in a way that’s intelligent” (Al-Sibai, 2025)

Generative AI reduces cognitive load – good and bad (Lee, 2025)

- May improve worker efficiency
- But can also inhibit critical engagement with work and potentially lead to
  - Long-term overreliance on tool
  - Diminished skill for problem-solving
- Knowledge workers engage in critical thinking to ensure quality of their work, i.e., verifying outputs against external sources
- When using GenAI tools, effort in critical thinking shifts from
  - Information gathering to information verification
  - Problem-solving to AI response integration
  - Task execution to task stewardship

Does ChatGPT impact cognition (Kosmyna, 2025)?

- College students (n=54) randomized and assigned to write essay
  - Brain-only
  - Search engine
  - LLM
- Later session assigned
  - Brain-only ➡️ LLM
  - LLM ➡️ Brain-only
- EEG monitoring showed differences in brain connectivity
  - Brain-only > search engine > LLM
- Essay grading found
  - ”Ownership” of essays strongest for brain-only > search engine > LLM
  - LLM users had difficulty accurately quoting own work immediately and 4 months later

How does this impact clinical and informatics practice going forward?

- Performance of models with complex medical cases is very impressive and likely will play a role in clinician work in the future
- Need roadmap to lessen cheating and over-reliance on GenAI by maintaining student-centered focus and integrity (Gallant, 2025)
- Current systems and research about them do not replicate real-world clinical practice
  - Medicine is rarely a zero-shot or one-shot activity, i.e., doctors have more chances to get it right
  - Well-curated clinical cases are rarely the norm in medical decision-making
  - An important part of clinical practice is knowing the right questions to ask and proper data to collect
  - Doctoring is also about supporting patients and their families in challenging times