# Results and Implications for Generative AI in Education

William Hersh, MD
Professor
Department of Medical Informatics & Clinical Epidemiology
School of Medicine
Oregon Health & Science University
Portland, OR, USA
https://www.ohsu.edu/informatics
Email: hersh@ohsu.edu
Web: http://www.billhersh.info/
Blog: https://informaticsprofessor.blogspot.com/
Twitter: @williamhersh

References

Benoit, J.R.A., 2023. ChatGPT for Clinical Vignette Generation, Revision, and Evaluation. https://doi.org/10.1101/2023.02.04.23285478

Cabral, S., Restrepo, D., Kanjee, Z., Wilson, P., Crowe, B., Abdulnour, R.-E., Rodman, A., 2024. Clinical Reasoning of a Generative Artificial Intelligence Model Compared With Physicians. JAMA Intern Med. https://doi.org/10.1001/jamainternmed.2024.0295

Chen, A., Chen, D.O., Tian, L., 2023. Benchmarking the symptom-checking capabilities of ChatGPT for a broad range of diseases. J Am Med Inform Assoc ocad245. https://doi.org/10.1093/jamia/ocad245

Cheng, L., Li, X., Bing, L., 2023. Is GPT-4 a Good Data Analyst?, in: Bouamor, H., Pino, J., Bali, K. (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023. Presented at the Findings 2023, Association for Computational Linguistics, Singapore, pp. 9496–9514. https://doi.org/10.18653/v1/2023.findings-emnlp.637

Choi, J.H., Monahan, A., Schwarcz, D., 2023. Lawyering in the Age of Artificial Intelligence. https://doi.org/10.2139/ssrn.4626276

de Vries, A., 2023. The growing energy footprint of artificial intelligence. Joule 7, 2191–2194. https://doi.org/10.1016/j.joule.2023.09.004

Denny, P., Prather, J., Becker, B.A., Finnie-Ansley, J., Hellas, A., Leinonen, J., Luxton-Reilly, A., Reeves, B.N., Santos, E.A., Sarsa, S., 2024. Computing Education in the Era of Generative AI. Commun. ACM 67, 56–67. https://doi.org/10.1145/3624720

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, Amy, Fan, A., Goyal, A., Hartshorn, A., Yang, Aobo, Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C.C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E.M., Radenovic, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G.L., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I.A., Kloumann, I., Misra, I., Evtimov, I., Copet, J., Lee, Jaewon, Geffert, J., Vranes, J., Park, Jason, Mahadeokar, J., Shah, J., van der Linde, J., Billock, J.,

Hong, J., Lee, Jenya, Fu, J., Chi, J., Huang, J., Liu, J., Wang, Jie, Yu, J., Bitton, J., Spisak, J., Park, Jongsoo, Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K.V., Upasani, K., Plawiak, K., Li, Ke, Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Rantala-Yeary, L., van der Maaten, L., Chen, Lawrence, Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., de Oliveira, L., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M.K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P.S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R.S., Stojnic, R., Raileanu, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, Rui, Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S.S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, Shun, Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, Xiaodong, Tan, X.E., Xie, X., Jia, X., Wang, Xuewei, Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Yuchen, Li, Yue, Mao, Y., Coudert, Z.D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Grattafiori, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Vaughan, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Franco, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., De Paola, B., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.-H., Cai, C., Tindal, C., Feichtenhofer, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Wyatt, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Ozgenel, F., Caggioni, F., Guzmán, F., Kanayet, F., Seide, F., Florez, G.M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Thattai, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Damlaj, I., Molybog, I., Tufanov, I., Veliche, I.-E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Asher, J., Gaya, J.-B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, Junjie, Wu, K., U, K.H., Saxena, K., Prasad, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, Keqian, Huang, Kun, Chawla, K., Lakhotia, K., Huang, Kyle, Chen, Lailin, Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Tsimpoukelli, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Seltzer, M.L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M.J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N.,

Egebo, N., Usunier, N., Laptev, N.P., Dong, N., Zhang, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Li, R., Hogan, R., Battey, R., Wang, Rocky, Maheswari, R., Howes, R., Rinott, R., Bondu, S.J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S.C., Shankar, S., Zhang, Shuqiang, Zhang, Shuqiang, Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Kohler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V.S., Mangla, V., Albiero, V., Ionescu, V., Poenaru, V., Mihailescu, V.T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wang, Xiaofang, Wu, Xiaojian, Wang, Xiaolan, Xia, X., Wu, Xilun, Gao, X., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Yenda, Zhang, Yilin, Zhang, Ying, Adi, Y., Nam, Y., Yu, Wang, Hao, Y., Qian, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., Zhao, Z., 2024. The Llama 3 Herd of Models. https://doi.org/10.48550/arXiv.2407.21783

Goh, E., Gallo, R., Hom, J., Strong, E., Weng, Y., Kerman, H., Cool, J., Kanjee, Z., Parsons, A.S., Ahuja, N., Horvitz, E., Yang, D., Milstein, A., Olson, A.P.J., Rodman, A., Chen, J.H., 2024. Influence of a Large Language Model on Diagnostic Reasoning: A Randomized Clinical Vignette Study. https://doi.org/10.1101/2024.03.12.24303785

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. Competencies and Curricula Across the Spectrum of Learners for Biomedical and Health Informatics. Stud Health Technol Inform 300, 93–107. https://doi.org/10.3233/SHTI220944

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

Hong, S., Lin, Y., Liu, Bang, Liu, Bangbang, Wu, B., Li, D., Chen, J., Zhang, J., Wang, J., Zhang, Li, Zhang, Lingyao, Yang, M., Zhuge, M., Guo, T., Zhou, T., Tao, W., Wang, W., Tang, X., Lu, X., Zheng, X., Liang, X., Fei, Y., Cheng, Y., Xu, Z., Wu, C., 2024. Data Interpreter: An LLM Agent For Data Science. https://doi.org/10.48550/arXiv.2402.18679

Johnson, M., 2024. Generative AI and CS Education. Commun. ACM 67, 23–24. https://doi.org/10.1145/3632523

Jones, N., 2024. 'In awe': scientists impressed by latest ChatGPT model o1. Nature. https://doi.org/10.1038/d41586-024-03169-9

Kanjee, Z., Crowe, B., Rodman, A., 2023. Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge. JAMA 330, 78–80. https://doi.org/10.1001/jama.2023.8288

Katz, U., Cohen, E., Shachar, E., Somer, J., Fink, A., Morse, E., Shreiber, B., Wolf, I., 2024. GPT versus Resident Physicians — A Benchmark Based on Official Board Scores. NEJM AI 0, AIdbp2300192. https://doi.org/10.1056/AIdbp2300192

Kim, A., Muhn, M., Nikolaev, V.V., 2024. Financial Statement Analysis with Large Language Models. https://doi.org/10.2139/ssrn.4835311

Kirkpatrick, K., 2023. The Carbon Footprint of Artificial Intelligence. Commun. ACM 66, 17–19. https://doi.org/10.1145/3603746

Kung, T.H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., Tseng, V., 2023. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health 2, e0000198. https://doi.org/10.1371/journal.pdig.0000198

Learning to Reason with LLMs [WWW Document], 2024. . OpenAI. URL https://openai.com/index/learning-to-reason-with-llms/ (accessed 10.7.24).

Levine, D.M., Tuwani, R., Kompa, B., Varma, A., Finlayson, S.G., Mehrotra, A., Beam, A., 2023. The Diagnostic and Triage Accuracy of the GPT-3 Artificial Intelligence Model. https://doi.org/10.1101/2023.01.30.23285067

Levkovich, I., Elyoseph, Z., 2023. Identifying depression and its determinants upon initiating treatment: ChatGPT versus primary care physicians. Fam Med Community Health 11, e002391. https://doi.org/10.1136/fmch-2023-002391

Mollick, E., 2024. Post-apocalyptic education [WWW Document]. One Useful Thing. URL https://www.oneusefulthing.org/p/post-apocalyptic-education (accessed 9.4.24).

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

Nori, H., Lee, Y.T., Zhang, S., Carignan, D., Edgar, R., Fusi, N., King, N., Larson, J., Li, Y., Liu, W., Luo, R., McKinney, S.M., Ness, R.O., Poon, H., Qin, T., Usuyama, N., White, C., Horvitz, E., 2023. Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine. https://doi.org/10.48550/arXiv.2311.16452

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).

Poldrack, R.A., Lu, T., Beguš, G., 2023. AI-assisted coding: Experiments with GPT-4. https://doi.org/10.48550/arXiv.2304.13187

Rao, A., Pang, M., Kim, J., Kamineni, M., Lie, W., Prasad, A.K., Landman, A., Dreyer, K., Succi, M.D., 2023. Assessing the Utility of ChatGPT Throughout the Entire Clinical Workflow: Development and Usability Study. J Med Internet Res 25, e48659. https://doi.org/10.2196/48659

Saab, K., Tu, T., Weng, W.-H., Tanno, R., Stutz, D., Wulczyn, E., Zhang, F., Strother, T., Park, C., Vedadi, E., Chaves, J.Z., Hu, S.-Y., Schaekermann, M., Kamath, A., Cheng, Y., Barrett, D.G.T., Cheung, C., Mustafa, B., Palepu, A., McDuff, D., Hou, L., Golany, T., Liu, L., Alayrac, J., Houlsby, N., Tomasev, N., Freyberg, J., Lau, C., Kemp, J., Lai, J., Azizi, S., Kanada, K., Man, S., Kulkarni, K., Sun, R., Shakeri, S., He, L., Caine, B., Webson, A., Latysheva, N., Johnson, M., Mansfield, P., Lu, J., Rivlin, E., Anderson, J., Green, B., Wong, R., Krause, J., Shlens, J., Dominowska, E., Eslami, S.M.A., Chou, K., Cui, C., Vinyals, O., Kavukcuoglu, K., Manyika, J., Dean, J., Hassabis, D., Matias, Y., Webster, D., Barral, J., Corrado, G., Semturs, C., Mahdavi, S.S., Gottweis, J., Karthikesalingam, A., Natarajan, V., 2024. Capabilities of Gemini Models in Medicine. https://doi.org/10.48550/arXiv.2404.18416

Shaffer, M., Wang, C.C.Y., 2024. Scaling Core Earnings Measurement with Large Language Models.

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

Tu, T., Palepu, A., Schaekermann, M., Saab, K., Freyberg, J., Tanno, R., Wang, A., Li, B., Amin, M., Tomasev, N., Azizi, S., Singhal, K., Cheng, Y., Hou, L., Webson, A., Kulkarni, K., Mahdavi, S.S., Semturs, C., Gottweis, J., Barral, J., Chou, K., Corrado, G.S., Matias, Y., Karthikesalingam, A., Natarajan, V., 2024. Towards Conversational Diagnostic AI. https://doi.org/10.48550/arXiv.2401.05654

# Results and Implications for Generative AI in Education

William Hersh
Professor
Department of Medical Informatics & Clinical Epidemiology
Oregon Health & Science University
Portland, OR, USA

1

## Goals for talk

- Provide overview of results for generative artificial intelligence (AI) in education and related tasks
- Present results of study comparing student with large language model (LLM) performance in an introductory biomedical and health informatics course
- Discuss implications of generative AI for student learning and assessment



2

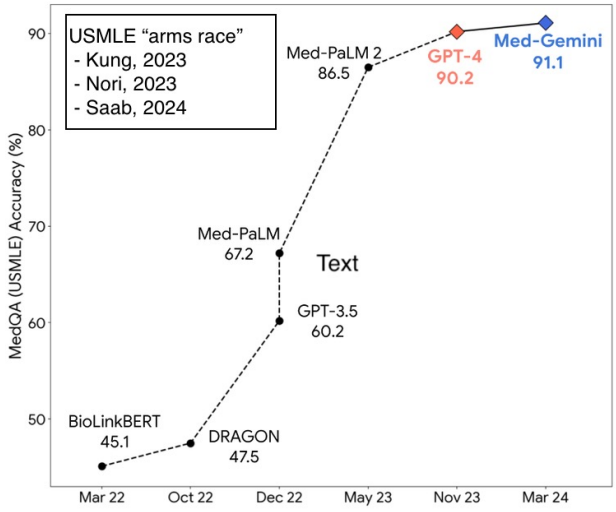
# Overview of results of generative AI in education and related tasks

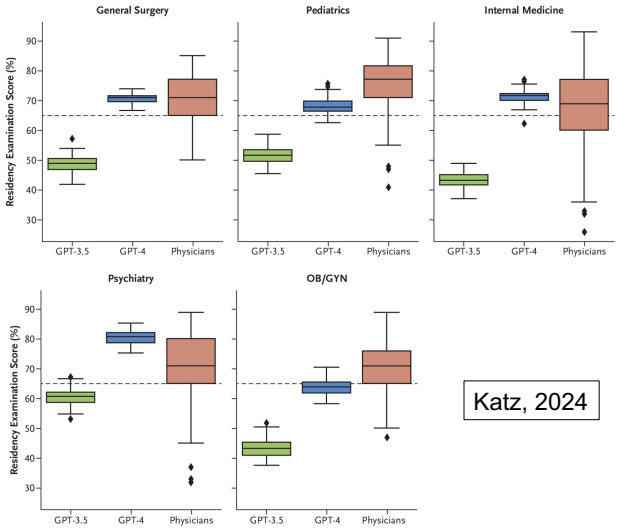
- LLMs perform well in many different types of knowledge assessments in biomedicine
  - Medical board exams
  - Graduate school bioscience exams
  - Objective structured clinical exams (OSCEs)
  - Answering clinical questions
  - Solving clinical cases
  - Conversational diagnostic dialogue
  - Clinical reasoning
- LLMs perform well in education in many other disciplines
  - High school standardized and AP exams
  - Computer science
  - Data science
  - Business
  - Law
  - PhD-level biology, chemistry, and physics





# Success of generative AI – medical board exams




Katz, 2024

# 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
- Two answer-sets flagged as plagiarism based on answer similarity
- Some model responses included detailed hallucinations





# Prompts and results

**GPT4-Simple Prompt Pattern:**

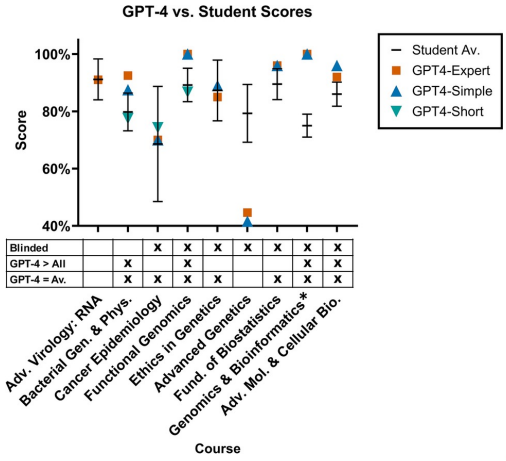
```
Please answer the following questions.
```

**GPT4-Expert Prompt Pattern:**

```
I am going to give you questions from an examination
in a graduate course in Cell Biology.
Please act as an expert in the field of Cell Biology.
Please answer each question as correctly as possible,
using technical or advanced language as necessary to
answer the question correctly.
```

**GPT4-Short Prompt Pattern (after GPT4-Expert response):**

```
Please shorten the last answer to approximately
sixty-five percent of the original length.
The shortened answer should be correct, clear,
and concise without any numeric lists.
```



# Other successes of LLMs (cont.)

- Solving clinical cases – comparable to but not better than expert humans (Levine, 2023; Kanjee, 2023; Rao, 2023; Benoit, 2023; Chen, 2023; Levkovich, 2023)
- In simulated (text-based) objective structured clinical exam (OSCE) format, Google's Articulate Medical Intelligence Explorer (AMIE) outperformed primary care physicians in text-based dialogue in history-taking, diagnostic accuracy, management reasoning, communication skills, and empathy (Tu, 2024)
- For 20 clinical cases, GPT-4 performed comparable to attending physicians and residents in diagnostic accuracy, correct clinical reasoning, and cannot-miss diagnosis inclusion (Cabral, 2024)
- In randomized vignette study of diagnostic reasoning, physicians with conventional information resources scored comparably with or without GPT-4 (76.3% vs. 73.7%, NS) but GPT-4 alone did better (92.1%, SS) (Goh, 2024)




7

# Success not limited to biomedicine – LLMs can

- Pass college entrance and AP exams (Dubey, 2024)
- Write computer programs (Poldrack 2024; Denny, 2024; Johnson, 2024)
- Create data science pipelines (Cheng, 2024; Hong, 2024)
- Predict company earnings (Kim, 2024; Shaffer, 2024)
- Write legal briefs (Choi, 2024)
- Outscore PhD students on "Google-proof" questions in biology, chemistry, and physics (OpenAI, 2024; Jones, 2024)

- Do we have a "homework apocalypse" (Mollick, 2023)?




8

## Concerns for LLMs

- Provide answers but not sources of knowledge (Hersh, 2024)
- Prone to hallucinations, confabulations, etc.
  - Dictionary.com 2023 word of year: hallucinate (Norlen, 2023)
- Speak with confidence and trustworthiness
- Impact on climate (Kirkpatrick, 2023)
  - One estimate is that electricity consumption of AI request is 10-fold more than Google search (de Vries, 2023)






## Goals of study (Hersh and Fultz Hollis, 2024) – working backwards through title

- 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?



# About the course

- Introductory overview about biomedical and health informatics (Hersh, 2007; Hersh, 2022)
  - Taught at OHSU for over three decades (Hersh, 2024)
  - Updated annually
- 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






# Offered in three versions

- BMI 510/610 – graduate-level course required for informatics students and elective for other (e.g., nursing, public health) students
  - Completed by 1683 students since 1996
- 10x10 (“ten by ten”) – continuing education course offered in partnership with American Medical Informatics Association (AMIA)
  - Completed by 3260 students since 2005
- MINF 705B/709A – elective course for medical students offered as two-week block or over academic quarter
  - Completed by 127 students since 2020



# Course outline – 2023

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

See (Hersh, 2024) for original course outline!






# 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



# Prompts

- MCQs
  - Each LLM 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.”
  - Followed by pasting in the MCQs one unit (10 questions) at a time exactly as they appeared in MCQ preview file in LMS
- Final exam
  - Each LLM 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.”
  - Followed by pasting in exam, which had 33 questions, separated into 8 sections with a one-sentence heading for each section, exactly as it appeared in LMS exam module






# 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

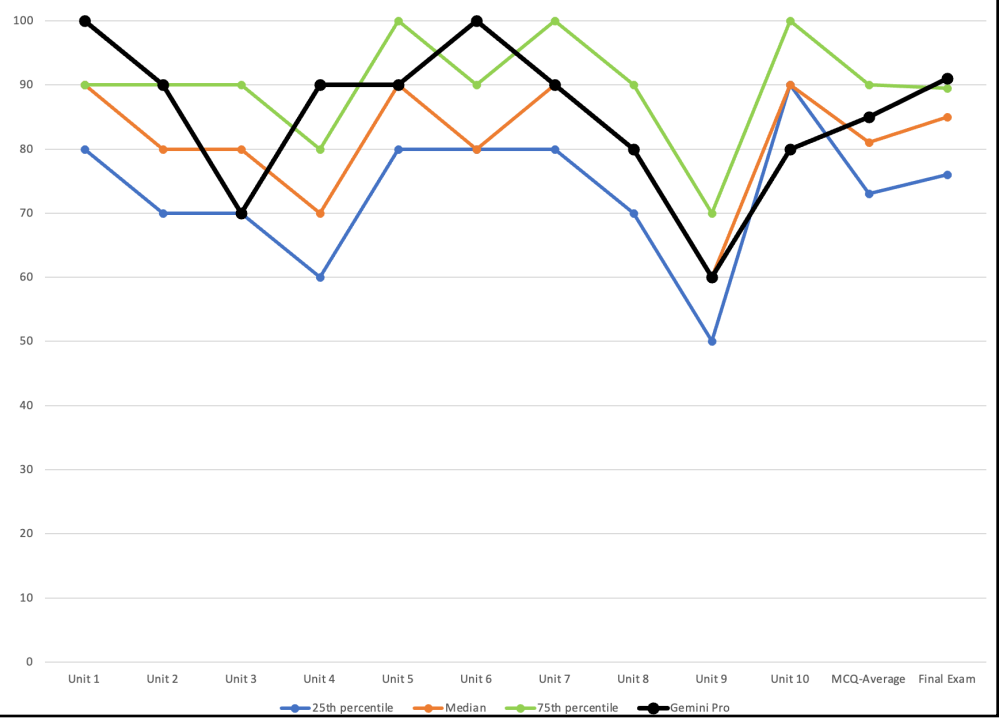
Which of the following is not a defined element of personal health information in the HIPAA Privacy Law?
a. Facial image
b. First and last name
c. Name of hospital where care is obtained
d. Personal email address
e. Twitter handle

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






17

# Results – detailed

| Students/LLMs | MCQ Unit Average | Final Exam | MCQ+Final Combined |
|---|---|---|---|
| | | | |
| Students – 25th percentile | 73 | 76 | 149 |
| Students – 50th percentile | 81 | 85 | 166 |
| Students – 75th percentile | 90 | 90 | 180 |
| | | | |
| ChatGPT Plus | 88 | 76 | 164 |
| Claude 3 Opus | 81 | 91 | 172 |
| CoPilot Bing-Precise | 88 | 85 | 173 |
| Gemini Pro | 85 | 91 | 176 |
| Llama 3.1 405B | 85 | 88 | 173 |
| Mistral-Large | 83 | 82 | 165 |



18

# 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 | | | | | | |





# Arguments with a peer reviewer

- Inferential statistics – why no p values?
- Why these LLMs used?
- Why use of Web interfaces and not APIs?
- More demographics on students?

## Limitations of study

- Single course
- Student LLM use before or in course unknown
- Biomedical and health informatics evolves rapidly
- (Major, true of all studies like this) LLMs challenging for reproducibility
  - Controlled by vendors
  - Constantly updated and changing
  - Little access to internals by most users






## Discussion

- First study of LLMs in a course like this in biomedical domain
- All LLMs scored at around 75$^{th}$ percentile for all students
- Raises issues about future of student learning expectations and assessments



# Now what?

**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.




23

# Next steps?

- What should be policy be for use of AI in our courses?
- How do we assess students in the era of generative AI?
- What is baseline knowledge for different disciplines?
- What else is important for generative AI and education?
- What might "post-apocalyptic education" look like (Mollick, 2024)?




24

# More information

- Paper
  - Hersh, W., Fultz Hollis, K. Results and implications for generative AI in a large introductory biomedical and health informatics course. *npj Digit Med.* 7, 247 (2024). https://doi.org/10.1038/s41746-024-01251-0
- OHSU News
  - https://news.ohsu.edu/2024/09/16/ohsu-researchers-test-chatgpt-other-ai-models-against-real-world-students
- Nature Research Communities
  - https://go.nature.com/3B1RLyO





# Thank you!

William Hersh, M.D.
Professor
Department of Medical Informatics & Clinical Epidemiology
Oregon Health & Science University
Portland, OR, USA
Email: hersh@ohsu.edu
Web: http://www.billhersh.info
Blog: https://informaticsprofessor.blogspot.com/
Also on
Facebook
LinkedIn
Twitter – @williamhersh