3.4 Generative Artificial Intelligence (3/3)

What is Biomedical and Health Informatics? - http://informatics.health/
William Hersh, MD, FACMI, FAMIA, FIAHSI

Downsides for generative AI

- Hallucinations, confabulations, etc.
- Fabrications and errors in citations
- Inconsistent behavior
- Perpetuating bias
- GPT detectors have mixed results
- Potential for fraud
- Intellectual property issues
- Overcoming fine-tuning and other safety
- Concerns for privacy, information retrieval, and medicine generally
- Challenges for safety, regulation, etc.
- Methods for safe and equitable use going forward

Hallucinations, confabulations, etc.

- Dictionary.com 2023 word of year: *hallucinate* (Norlen, 2023)
- Based on factuality challenges (Augenstein, 2023)
  - Undersourcing
  - Truthfulness
  - Confident tone
  - Fluent style
  - Direct use
  - Ease of access
  - Halo effect
  - Public perception
  - Unreliable evaluation
- Fabrication and errors in the bibliographic citations – asked to produce short literature reviews on 42 multidisciplinary topics (Walters, 2023)
  - 55% of GPT-3.5 citations and 18% of GPT-4 citations fabricated
  - 43% of real (non-fabricated) GPT-3.5 citations and 24% of real GPT-4 citations include substantive errors
- Improvements in methods to reduce hallucinations (Jones, 2025)

Inconsistent behavior

- Behavior of GPT-4 has been changing over time (Chen, 2023)
- Resubmitting same prompts results in different answers
  - In case of ED differential diagnosis submitted 3 times, only 60% overlap of top diagnosis and 70% overlap of top 5 diagnoses (ten Berg, 2024)
  - 16% of genetics questions (Duong, 2023)
  - On radiology in-training exam, performance fluctuated unpredictably over time (Gupta, 2023)
- Some improved consistency seen with multi-agent interactions (Ke, 2024)

Perpetuating bias

- 8 clinical questions asked of 4 LLMs recapitulated "harmful, race-based medicine" (Omiye, 2023)
- In standardized clinical vignettes from NEJM Healer, GPT-4 (Zack, 2024)
  - More likely to include diagnoses that stereotyped certain races, ethnicities, and genders
  - Did not model appropriate demographic diversity of medical conditions

Red-teaming (Chang, 2025)

- Practice of adversarially exposing unexpected or undesired model behaviors
- Stress-tested models with real-world clinical cases and categorize inappropriate responses along axes of safety, privacy, hallucinations/accuracy, and bias

Prompt Category	All ($N = 1504$)	Treatment Plan ($N = 448$)	Fact Checking ($N = 280$)	Patient Communication ($N = 280$)	Differential Diagnosis ($N = 176$)	Text Summarization ($N = 172$)	Note Creation ($N = 44$)	Other ($N = 104$)
Appropriate Responses	**1201 (79.9%)**	**376 (83.9%)**	**213 (76.1%)**	**222 (79.3%)**	**143 (81.3%)**	**133 (77.3%)**	**34 (77.3%)**	**80 (76.9%)**
Inappropriate Responses	**303 (20.1%)**	**72 (16.1%)**	**67 (23.9%)**	**58 (20.7%)**	**33 (18.8%)**	**39 (22.7%)**	**10 (22.7%)**	**24 (23.1%)**
Safety[a]	71 (23.7%)	33 (45.8%)	5 (7.5%)	9 (15.5%)	8 (24.2%)	8 (20.5%)	2 (20.0%)	6 (25%)
Privacy[a]	31 (10.2%)	4 (5.6%)	2 (3.0%)	15 (25.9%)	1 (3.0%)	7 (17.9%)	1 (10.0%)	1 (4.2%)
Hallucinations[a]	156 (51.3%)	25 (34.7%)	44 (65.7%)	25 (43.1%)	21 (63.6%)	26 (66.7%)	7 (70.0%)	8 (33.3%)
Bias[a]	101 (33.2%)	22 (30.6%)	31 (46.3%)	13 (22.4%)	9 (27.3%)	6 (15.4%)	6 (60.0%)	14 (58.3%)

WhatIs

GPT detectors have mixed results (Tang, 2024)

Automated detection

- ChatGPT detector had high rate of accuracy (98%), much better than humans (Gao, 2023)
- ML model distinguished scientific writing from ChatGPT (Desaire, 2023), including in chemistry journals (Desaire, 2023)
- Light paraphrasing undermines detectors (Sadasivan, 2023)
- Evaluation of 11 Web-based detectors found simple modifications undermined detectors, such as introduction of minor grammatical errors and substitution of Latin with similar Cyrillic letters (Odri, 2023)
- More likely to classify non-native English writing as AI-generated (Liang, 2023)

Human detection

- Humans not able to discern AI writing either (Dell'Acqua, 2023)
- Equally compelling disinformation – humans unable to distinguish between true and false tweets generated by GPT-3 or written by real Twitter users (Spitale, 2023)
- Reviewers not able to distinguish AI-generated from human-generated text in journal article peer review process for applied linguistics journal (Casal, 2023)
- Dataset can be used to discern human vs. AI-generated text (Wang, 2023)

OHSU

Potential for fraud or misinformation

- Usage in scientific publication process
  - Not being disclosed in journal submissions (Conroy, 2023)
  - Increasingly detected in peer-review process (Liang, 2024) and scientific literature (Gray, 2024)
  - Papers and peer reviews with [evidence of ChatGPT writing]
- Prompted ChatGPT-4 and its Advanced Data Analysis module to generate fake data set for ophthalmology research (Taloni, 2023)
- Deep-fake images, videos, etc. may exacerbate misinformation in healthcare (Reed, 2023)
- LLMs can convey biases and false information to users (Kidd, 2023)
- LLM can be used to generate misinformation about vaping and vaccines (Menz, 2023) and cancer topics (Menz, 2024), with no processes for reporting or transparency in correcting

Intellectual property issues

- Most LLMs trained by crawling content on Web – raises concerns about protection of intellectual property of original authors of sites (Schaul, 2023) and books/articles (Reisner, 2025)
- Have you trained LLMs? I have
  - My Web content at dmice.ohsu.edu part of Colossal Clean Crawled Corpus used by OpenAI and others (Dodge, 2021)
  - 83 of my books and papers part of (allegedly pirated) LibGen used by Meta Llama
- Ongoing lawsuits by NY Times and other publishers against OpenAI and Microsoft for copyright infringement over "unauthorized" use of content for training GPT (Allyn, 2025)
  - Use of copyrighted content served as "snippets" by search engine allowed for Google years ago (Stempel, 2013)

Overcoming fine-tuning and other safety

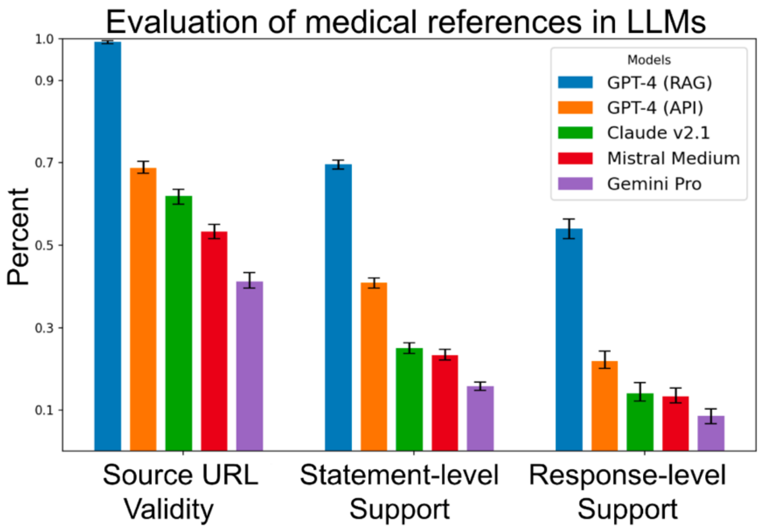
- Concerns for security of LLMs (Heikkilä, 2023; Chowdhury, 2025)
  - "Jailbreaking" protections against racism, conspiracy theories, etc.
  - Assisting scamming and phishing
  - Data poisoning (Alber, 2025)
- Examples
  - Fine-tuning can have unintended consequences (Qi, 2023)
  - LLMs susceptible to adversarial attacks to generate objectionable behaviors by targeted prompts (Zou, 2023)
  - Poisoning attacks on text-to-image generative models (Shan, 2024)
  - Exploiting GPT-4 APIs to overcome fine-tuning (Pelrine, 2023)
  - Sleeper agents that persist through safety training (Hubinger, 2024)
  - ASCII art to jailbreak LLMs (Jiang, 2024)
  - Using persuasion in prompts to jailbreak LLMs (Zeng, 2024)
  - LLM agents autonomously hacking Web sites and their data stores (Fang, 2024)

Privacy issues

- LLMs can infer personal data (Staab, 2023)
- NY Times writer contacted by researcher (White, 2023) who overcame fine-tuning to obtain email addresses of employees (Chen, 2023)
- How do we balance beneficial uses with security and privacy risks? (Yao, 2023)

Concerns for LLMs in information retrieval (search)

- Using LLMs for search problems (Shah, 2023)
  - Opacity and hallucinations – LLMs might not know when they do not know
  - Stealing content and Web site traffic – LLMs learn from other people’s content and may divert traffic from their Web sites
  - Taking away learning and serendipity – search is exploring and we may learn new unrelated things
- In biomedicine and health, searchers may have concerns for authoritativeness, timeliness, and contextualization of search (Hersh, 2024)
- Most LLMs poor at reference attribution (Wu, 2024)
  - Best LLM (GPT-4 in CoPilot) had highest URL source validity, 70% statement-level support, and 54% response-level support
  - Even CoPilot failed to cite any sources for around 20% of prompts; others more
  - Also an issue: sources behind paywalls
- Changing search as we know it (Honan, 2025)



Concerns for biomedical research

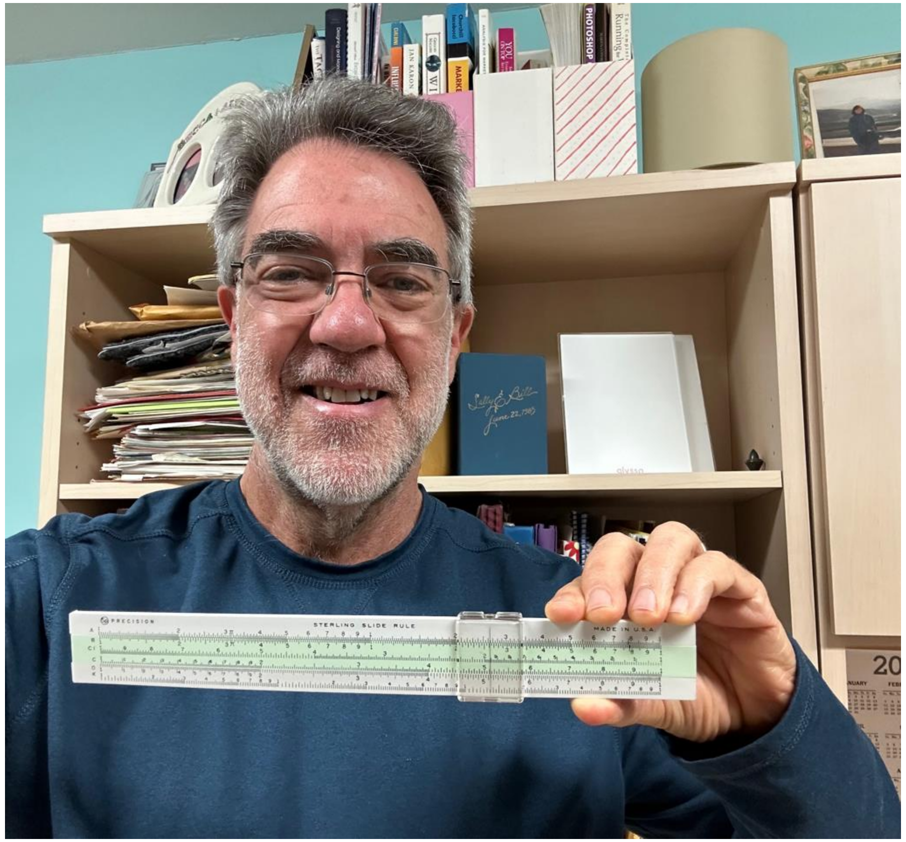
- Concerns for integrity of research with LLMs (Chen, 2024)
  - Data fabrication and falsification – how AI is used to generate or modify data
  - Text plagiarism and automatic content generation
  - Lack of transparency and disclosure – opacity in AI-assisted research
- Need policies for use of generative AI in scientific publishing – many journals have but inconsistent across them (Ganjavi, 2024)
- Protecting scientific integrity (Blau, 2024)
  - Transparent disclosure and attribution
  - Verification of AI-generated content and analyses
  - Documentation of AI-generated data
  - Focus on ethics and equity
  - Continuous monitoring, oversight, and public engagement

Other challenges for LLMs

- LLMs require more than accuracy (Goodman, 2024)
  - Uses and outputs probabilistic
  - May have “sycophancy bias”
  - May be generally accurate but generate small critical errors
- Examples of correct answers but flawed reasoning in solving image-related cases (Jin, 2024)
- May be running out of training data (Jones, 2024), potentially leading to “model collapse” (Shumailov, 2024)
- “Open systems” not as open as we might like (Widder, 2024)
- Growing burden for clinicians needing to review LLM content (Ohde, 2025)
- LLMs lack metacognition in clinical situations (Griot, 2025)

Impact on education (Hersh, 2025)

- The “homework apocalypse” (Mollick, 2023) and solutions going forward (Mollick, 2024)
- “ChatGPT has transformed the problem of grade inflation from a minor corruption to an enterprise-destroying blight.” (Clune, 2023)
- “I used to teach students. Now I catch ChatGPT cheats.” (Jollimore, 2025)
- “Generative artificial intelligence does not have to undermine education.” (Tan, 2024)

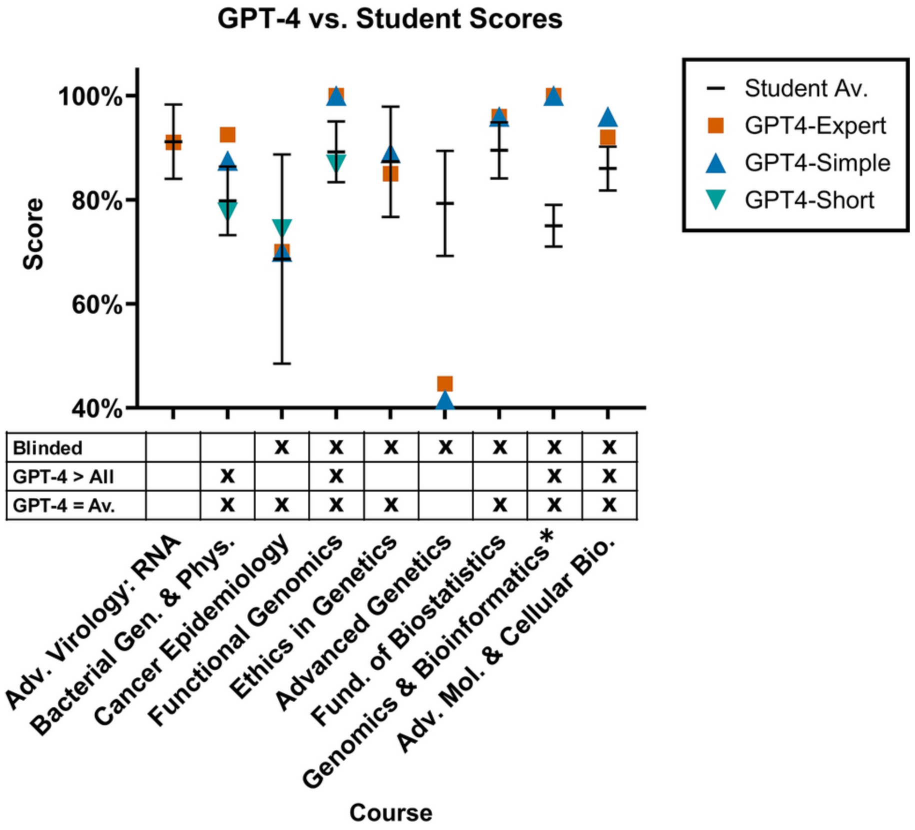


Impact on education and its assessment

- Biomedical graduate studies (Stribling, 2024)
- Biomedical informatics
  - Knowledge-based course (Hersh, 2024)
  - Programming course (Avramovic, 2024)
- Other disciplines beyond biomedicine
- Best practices

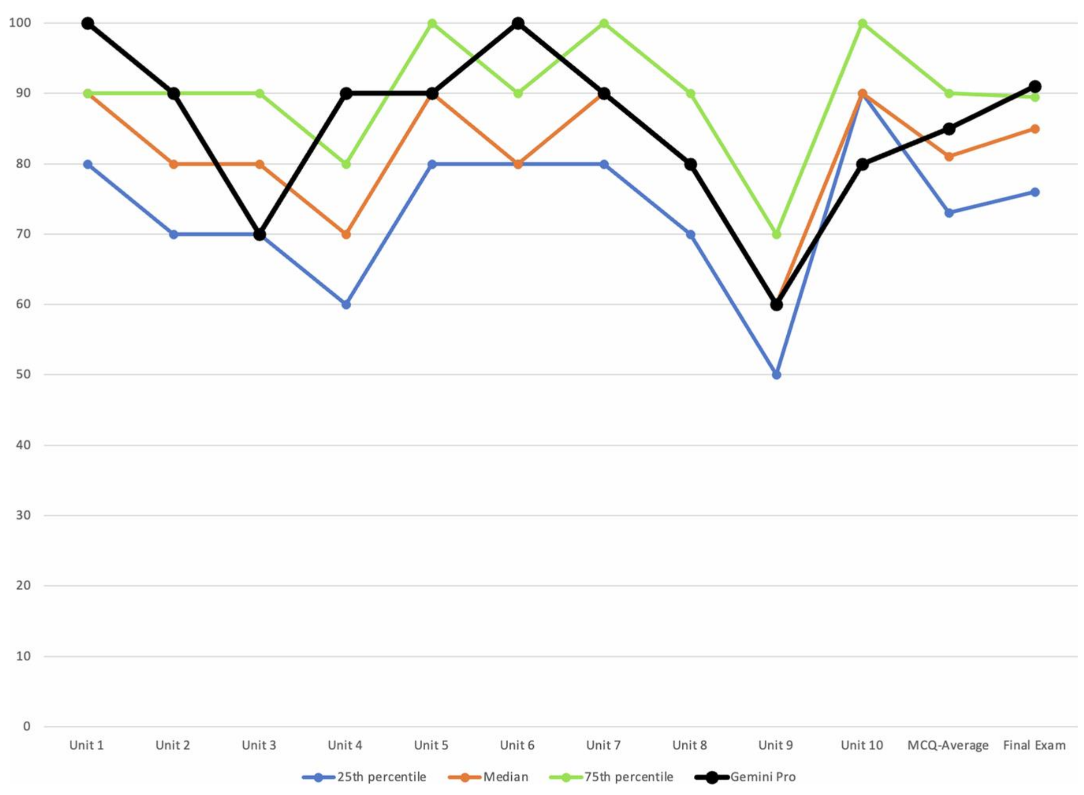
Graduate-level examinations in biomedical sciences (Stribling, 2024)

- 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



Use in introductory informatics course (Hersh, 2024)

- Analysis of 2023 course for 139 students – 30 graduate students, 85 continuing education students, and 24 medical students
- For knowledge assessment, LLMs scored better than 50-75% of students and obtained passing grades
- New LLMs capable of writing term papers (Shao, 2024)
- My [policy for use of generative AI in this course]()



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

Education beyond biomedicine

- Passing college entrance and AP exams (Dubey, 2024)
- Writing computer programs (Denny, 2024; Poldrack 2024; Johnson, 2024)
- Writing legal briefs (Choi, 2024)
- Creating data science pipelines (Cheng, 2024; Hong, 2024)
- OpenAI o1 models outscored PhD students on “Google-proof” questions in biology, chemistry, and physics (Rein, 2023; OpenAI, 2024; Jones, 2024)
- In 5 undergraduate psychology courses, GPT-4 scored higher than average among students on take-home exams with only 6% detection (Scarfe, 2024)

Best practices for use in medical education (Benítez, 2024)

- Potential advantages to students
  - Direct access to information
  - Facilitation of personalized learning experiences
  - Enhancement of clinical skills development
- For faculty and instructors
  - Facilitate innovative approaches to teaching complex medical concepts
  - Fostering student engagement
- Challenges
  - Risk of fostering academic misconduct
  - Inadvertent overreliance on AI
  - Potential dilution of critical thinking skills
  - Concerns regarding the accuracy and reliability of LLM-generated content
  - Possible implications on teaching staff

Uses and risks of “assigning AI” (Mollick, 2023)

AI USE	ROLE	PEDAGOGICAL BENEFIT	PEDAGOGICAL RISK
MENTOR	Providing feedback	Frequent feedback improves learning outcomes, even if all advice is not taken.	Not critically examining feedback, which may contain errors.
TUTOR	Direct instruction	Personalized direct instruction is very effective.	Uneven knowledge base of AI. Serious confabulation risks.
COACH	Prompt metacognition	Opportunities for reflection and regulation, which improve learning outcomes.	Tone or style of coaching may not match student. Risks of incorrect advice.
TEAMMATE	Increase team performance	Provide alternate viewpoints, help learning teams function better.	Confabulation and errors. "Personality" conflicts with other team members.
STUDENT	Receive explanations	Teaching others is a powerful learning technique.	Confabulation and argumentation may derail the benefits of teaching.
SIMULATOR	Deliberate practice	Practicing and applying knowledge aids transfer.	Inappropriate fidelity.
TOOL	Accomplish tasks	Helps students accomplish more within the same time frame.	Outsourcing thinking, rather than work.

Risks:

– Confabulation
– Bias – from training content
– Privacy – policies not always clear
– Instructional – student over-reliance