Generative AI in Biomedicine and Health

What is Biomedical and Health Informatics? - http://informatics.health/
William Hersh

Oregon Health & Science University

1

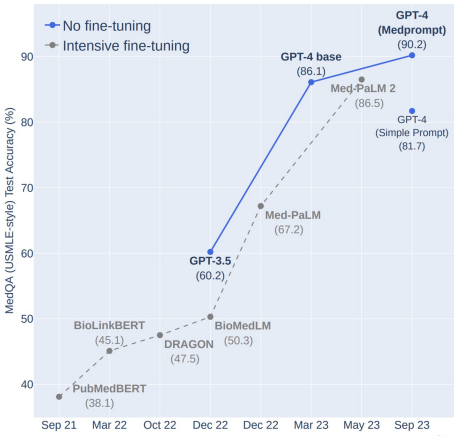
Uses and research with generative AI

- Board examination questions
- Answering questions
- Solving clinical cases
- Patient tasks
- Other tasks
- Artificial general intelligence

2

Board examination questions

- USMLE “arms race” on MedQA data set (Jin, 2021)
  - Original ChatGPT first to achieve passing-level score (60.2%) (Kung, 2023)
  - GPT-4 without any specialized prompt crafting exceeded passing score on USMLE by over 20% and outperformed GPT-3.5 and models specifically fine-tuned on medical knowledge (e.g., Med-PaLM) (Nori, 2023)
  - GPT-4 did well even on “soft skills” (e.g., communication skills, ethics, empathy, and professionalism) questions (Brin, 2023)
  - Best published score yet (90.2%) used GPT-4 with combination of several prompting strategies (Nori, 2023)



Board examination questions (cont.)

- ChatGPT-3.5 answered 74% of 254 questions correctly from clinical informatics board review book, above 60% threshold for passing (Kumah-Crystal, 2023)
- ChatGPT-3.5 answered 45% of 936 questions correctly from neonatal board review book, unlikely to pass exam (Beam, 2023)
  - Better on knowledge recall and basic clinical reasoning than multilogical questions
- GPT-4 answered 81% of 150 questions correctly on questions similar to radiology board exam (Bhayana, 2023)
  - Better than GPT-3.5, which scored better on lower-level questions than more integrative ones (Bhayana, 2023)
- ChatGPT-3.5 and ChatGPT-4 achieved scores of 73.4% and 83.4%, respectively, relative to user average of 72.8% on a 500-question neurosurgical written board examination (Ali, 2023)
- On ACR radiation oncology in-training (TXIT) exam and Red Journal Gray Zone cases, ChatGPT-3.5 and ChatGPT-4 achieved scores of 62.1% and 78.8% respectively, varying by different areas (Huang, 2023)
- GPT-4 scored passing level and better than average humans on neurology question bank (Schubert, 2023)

Other medical examinations

- Mixed results of passing and not; generally best performance with GPT-4
- ChatGPT scored better (77.2%) than humans (73.7%) on virtual objective structured clinical examination (OSCE) in obstetrics and gynecology and took one-quarter of time (Li, 2023)
- ChatGPT achieved likely passing score on European Exam in Core Cardiology (EECC), final exam for completion of specialty training in Cardiology in many countries (Skalidis, 2023)
- Ophthalmology Knowledge Assessment Program (OKAP)
  - ChatGPT-3.5 achieved 49.2-59.4% correct (Antaki, 2023)
  - ChatGPT-4 achieved 81% correct (Teebagy, 2023)
- Nephrology Self-Assessment Program (nephSAP) (Wu, 2024)
  - ChatGPT-4 scored 73.3%, much higher than other LLMs

Answering clinical questions

- Many studies on many clinical areas using many approaches – successes demonstrate how well LLMs can perform on answering clinical questions but negative studies show there is not always success
- On questions of radiation oncology physics, ChatGPT-4 scored better than other LLMs (BARD, BLOOMZ, and GPT-3.5) and humans, but not as well as team of human experts (Holmes, 2023)
- Using ChatGPT for concordance with National Comprehensive Cancer Network treatment guidelines for breast, prostate, and lung cancer (Chen, 2023)
  - Concordant 61.9% of time overall
  - 34.3% of outputs recommended 1 or more nonconcordant treatments
  - Responses hallucinated (i.e., not part of any recommended treatment) in 13 of 104 (12.5%) outputs
- On 284 physician-developed questions, ChatGPT-4 had highly accurate and complete answers, better than ChatGPT-3.5 (Goodman, 2023)

Answering clinical questions (cont.)

- ChatGPT-3.5 answered 12 of 38 questions (31.6%) on actinic keratosis (AK) with accurate, current, and complete response (Lent, 2023)
  - Performed best for questions on patient education, including pathogenesis of AK and potential risk factors, but did less well with diagnosis and treatment
  - Major deficits seen in grading AK, providing up-to-date treatment guidance, and asserting incorrect information with unwarranted confidence
- Answering 85 multiple-choice questions about human genetics (Duong, 2023)
  - ChatGPT 68.2% accurate, compared to 66.6% accuracy for humans
  - Both ChatGPT and humans performed better on memorization-type questions than on critical thinking questions
  - When asked same question multiple times, ChatGPT provided different answers 16% of time, including for both initially correct and incorrect answers, and gave plausible explanations for both correct and incorrect answers

Answering clinical questions (cont.)

- ChatGPT-4 more accurate than clinicians in determining pretest and posttest probability after negative test result in 5 cases but did not perform as well after positive test results (Rodman, 2023)
- For 66 questions submitted to hospital consultation service, evaluated by 12 physicians who voted to agree, disagree, or be unable to assess concordance with consult service recommendations (Dash, 2023)
  - For GPT-3.5, 37 questions had majority responses – 8 questions concordant, 20 discordant, and 9 unable to be assessed
  - For GPT-4, 37 questions had majority responses – 13 questions concordant, 15 discordant, and 3 unable to be assessed
  - Responses from both LLMs largely devoid of overt harm, but less than 20% of responses overall agreed with answer from consultation service
  - Some responses contained hallucinated references

Solving clinical cases with ChatGPT

- Use of 45-48 vignettes previously developed to assess symptom-checkers
  - https://scholar.harvard.edu/mehrotra/symptom-checkers
  - Earlier studies found physicians had 72% accuracy on vignettes (Semigran, 2016)
- Assessed with ChatGPT-3.5 for first-pass diagnostic and triage decision accuracy
  - Achieved 75.6% first-pass diagnostic accuracy and 57.8% triage accuracy (Benoit, 2023)
  - Also useful for generating new vignettes for high and low health literacy levels
- Assessing diagnostic and triage accuracy with ChatGPT-3.5 (Levine, 2023)
  - Correct diagnosis in top 3 for 88% of cases, compared to 54% for lay individuals and 96% for physicians
  - Triaged 71% correct, similar to lay individuals (74%), both worse than physicians (91%)

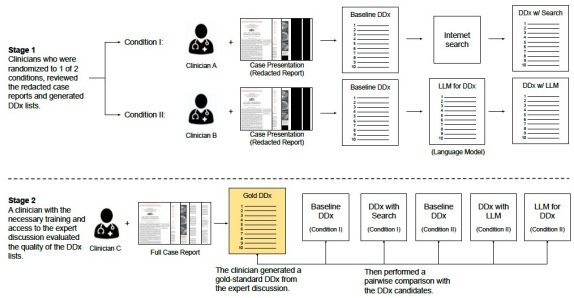
Solving clinical cases (cont.)

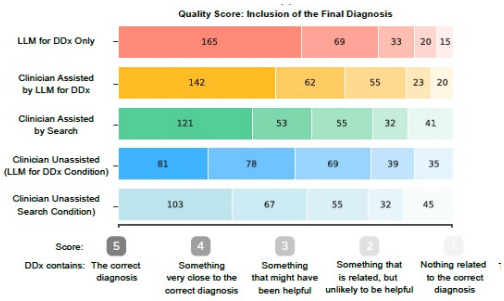
- For "challenging" New England Journal of Medicine (NEJM) clinicopathologic conferences
  - GPT-4 provided correct diagnosis within differential diagnosis in 64% of 70 cases and as top diagnosis in 39% (Kanjee, 2023)
  - GPT-4 correct for 57% of 38 cases, better than almost all online readers who answered (Eriksen, 2023)
    - Performance comparable for cases newer and older than September 2021 training of GPT-4
- For 36 MSD Clinical Manual clinical vignettes (Rao, 2023)
  - https://www.merckmanuals.com/professional/pages-with-widgets/case-studies
  - Overall correctness on all questions for all cases – 71.7%
  - Highest performance for final diagnosis – 76.9%
  - Lowest performance for generating initial differential diagnosis – 60.3%
  - Overall accuracy lower for diagnostic and management questions than for diagnosis questions
  - No variation by age, gender, or acuity of patient

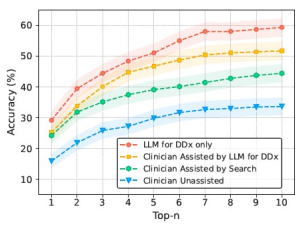
Solving clinical cases (cont.)

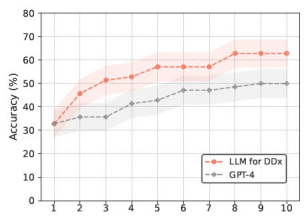
- User study of Google Med-PaLM 2 optimized for diagnostic clinical reasoning applied to 302 NEJM clinicopathologic conferences (McDuff, 2023)
- Generalist physicians given version of cases redacted for diagnostic testing and final diagnosis, asked to generate differential diagnosis (DDx) when randomized to two conditions – access to search vs. access to output from Med-PaLM 2
- Specialist physicians with access to gold standard evaluated DDx lists, evaluated DDx lists for inclusion of final diagnosis, comprehensiveness of DDx, and appropriateness of DDx
- Overall best DDx from LLM only, followed by generalist physicians with Med-PALM2, generalist physicians with search, and unassisted generalist physicians (next slide)
- Google LLM exceeded performance of GPT-4

User study of LLM for DDx (McDuff, 2023)









Solving clinical cases (cont.)

- Retrospective review of notes in Dutch ED for generating differential diagnosis (ten Berg, 2024)
  - For History & Physical, correct inclusion of diagnosis in top 5 of differential was 83% for physicians, 77% for ChatGPT-3.5, and 87% for ChatGPT-4
  - When lab data included, physicians accuracy increased to 87%, ChatGPT-3.5 to 97%, and ChatGPT-4 remained at 87%
  - Physicians chose correct leading diagnosis in 60% of cases, compared to ChatGPT-3.5 (37%) and ChatGPT-4 (53%)
  - With laboratory results, physicians chose correct leading diagnosis in 53% of cases, comparable to the accuracy of ChatGPT-3.5 (60%) and ChatGPT-4 (53%)
  - Submitting identical query to ChatGPT-3.5 or 4 3 different times had same leading diagnosis only 60% of time and overlap of all differential only 70% of time
- ChatGPT-4 aligned well with accepted guidelines for managing mild and severe depression, without showing the gender or socioeconomic biases observed among primary care physicians (Levkovich, 2023)

Solving clinical cases (cont.)

- For 194 diseases in Mayo Clinic Symptom Checker, ChatGPT-4 achieved 78.8% accuracy in making diagnosis, varying by clinical specialty (Chen, 2023)
- 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)

Answering patient/consumer questions

- ChatGPT-3.5 answered 21 of 25 questions about cardiovascular disease prevention deemed acceptable by cardiology clinicians for patient-facing information platform and as AI-generated draft responses to questions sent by patients (Sarraju, 2023)
- ChatGPT-3.5 provided evidence-based answers to public health questions, although primarily offered advice rather than referrals to potentially valuable resources (Ayers, 2023)
- ChatGPT-4 responses to patient questions posted to public social media forum rated higher quality and more empathetic (Ayers, 2023)

15

Answering patient/consumer questions (cont.)

- For 200 eye care questions from online advice forum, ChatGPT-3.5 generated appropriate responses not significantly different from ophthalmologist responses in terms of incorrect information, likelihood of harm, extent of harm, or deviation from ophthalmologist community standards (Bernstein, 2023)
- Different prompts impact correctness – in health misinformation dataset, results worse when evidence presented along with question (Koopman, 2023)
- Anecdote – ChatGPT solved case where 17 doctors over 3 years could not diagnose chronic pain in a child from spina bifida (Holohan, 2023; Venkataraman, 2023)

16

Answering patient/consumer questions (cont.)

- 4 AI chatbots – ChatGPT version 3.5 (OpenAI), Perplexity (Perplexity.AI), Chatsonic (Writesonic), and Bing AI (Microsoft)
  - For consumer cancer-related search queries, generally produced accurate information but responses not readily actionable and written at a college-reading level (Pan, 2023)
  - For urologic malignancies, produce information generally accurate and of moderately high quality but responses fairly difficult to read, moderately hard to understand, and lack clear instructions for users to act upon (Musheyev, 2023)
- For questions on safety of COVID-19 vaccines, ChatGPT default responses incomplete but generally satisfactory (Salas, 2023)

17

Generating patient letters and messages

- ChatGPT-3.5 wrote patient clinic letters with high level of correctness and measure of “humanness” (Ali, 2023)
- For 36 risks, benefits, and alternatives (RBAs) for common surgical procedures, ChatGPT-3.5 generated more readable, complete, and accurate consent documentation than surgeons (Decker, 2023)
- ChatGPT-3.5 asked to generate simplified radiology reports found to be factually correct, complete, and not potentially harmful to patient but with instances of incorrect statements, missed relevant medical information, and potentially harmful passages (Jeblick, 2023)
- New model (CLAIR) based on fine-tuning LLaMA-65B model generated patient portal messages deemed positive for responsiveness, empathy, and accuracy and neutral for usefulness (Liu, 2023)
- Pilot study of clinical usage of draft letters found about 20% utilization for task, with significant reductions in burden and burnout score derivatives but no change in time taken (Garcia, 2024)

18

Ambient dictation/virtual scribes

- Drafting clinical notes based on "listening in" to patient-clinician encounter
- Several commercial products, e.g., Nuance DAX and Abridge, and taken up by EHR vendors, e.g., Epic
- Early results show systems produce high-quality documentation, physician satisfaction, and reduced time both in clinic and after hours (Haberle, 2024; Tierney, 2024; ScribeAmerica, 2024)
- GPT-4 used to generate SOAP notes based on simulated patient-provider transcripts less successful (Kernberg, 2023)

Document summarization

- Hospital discharge summaries
  - Using clinical guidelines and notes, achieved accuracy of 81% (Ellershaw, 2024)
  - Transforming for patients rated patient-friendly but 44% not entirely complete and 18% found safety concerns for incomplete or inaccurate information (Zaretsky, 2024)
- Journal articles
  - GPT-4 feedback on scientific papers (Liang, 2023)
    - For PDFs of papers, found to have overlap comparable to between humans; higher for poorer-quality papers
    - Over half (57.4%) of authors found generated feedback helpful/very helpful and 82.4% found it more beneficial than feedback from at least some human reviewers
  - Summaries of 140 evidence-based journal abstracts generated by ChatGPT 70% shorter than mean abstract length and found to have high quality, high accuracy, and low bias (Hake, 2024)

ChatGPT on other tasks

- Clinical decision support alerts
  - For 7 alerts, ChatGPT generated suggestions unique from humans deemed to be valuable (Liu, 2023)
  - GPT-4 summarized clinician reasons for overriding alerts (Liu, 2024)
- GPT-3.5 and GPT-4 fared poorly in generating International Classification of Diseases (ICD)-9/10 codes (Soroush, 2023)
- For postoperative patient instructions, ChatGPT-generated instructions scored lower in understandability, actionability, and procedure-specific content than Google Search and institution-specific instructions (Ayoub, 2023)
- Information extraction from EHR
  - Rare disease identification and phenotype extraction (Shyr, 2024)
  - Identifying social determinants of health (Guevara, 2024)

Closing the loop? – LLMs for predictive AI

- Combining PaLM with radiology reports and an image encoder enabled zero-shot detection of five CXR findings – atelectasis, cardiomegaly, consolidation, pleural effusion, and pulmonary edema (Xu, 2023)
- Uniform transformer-based model outperformed image-only and non-unified models for prediction of adverse events in pulmonary disease (Zhou, 2023)
- For CXRs in ED, prior CXR plus report with LLM produced similar clinical accuracy and textual quality to on-site radiologist reports while providing higher textual quality than teleradiologist reports (Huang, 2023)
- LLaVA-Med uses image-report pairing to answer questions about images (Li, 2023)
- GPT-4 achieved accuracy of predicting cardiovascular disease comparable to Framingham model with data from UK Biobank and Korean Genome and Epidemiology Study (Han, 2023)

Some results from outside healthcare

- ChatGPT-3.5 and 4 outperformed average but not best humans in Alternative Uses Task, a creative divergent thinking task (Koivisto, 2023)
- ChatGPT-4 exceeded other LLMs at variety of general human language tasks but performs less well on reasoning tasks and is prone to hallucinations (Bang, 2023)
- GPT-4 created computer code for complex tasks but still required humans to ensure validity and accuracy (Poldrack, 2023)
- Assignment of occupation-specific, incentivized writing tasks to 453 college-educated professionals found 40% decreased time and 18% improved quality for half using ChatGPT (Noy, 2023)
- At global management consulting firm, consultants randomized to using ChatGPT-4 were (Dell'Acqua, 2023)
  - Significantly more productive – completed 12.2% more tasks on average, and completed task 25.1% more quickly)
  - Produced significantly higher quality results – more than 40% higher quality compared to control group
  - Noted to be part of "jagged technological frontier" where some tasks easily done by AI and others not, such as combining qualitative and quantitative data

Toward artificial general intelligence?

- GPT-4 can solve novel and difficult tasks that span mathematics, coding, vision, medicine, law, psychology, and more, without needing any special prompting (Bubeck, 2023)
  - Shows "sparks of artificial general intelligence"
  - "Strikingly close to human-level performance" that often vastly surpasses prior models such as ChatGPT-3.5
- Transformer system using meta-learning for compositionality (MLC) achieved human-like systematic generalization, including learning from mistakes (Lake, 2023)
- GPT-3 performed as well as humans in reasoning by analogy (Webb, 2023)
- GPT-4 performed worse than humans in abstraction and analogy (Moskvichev, 2023) and abstraction and reasoning (Mitchell, 2023) problems