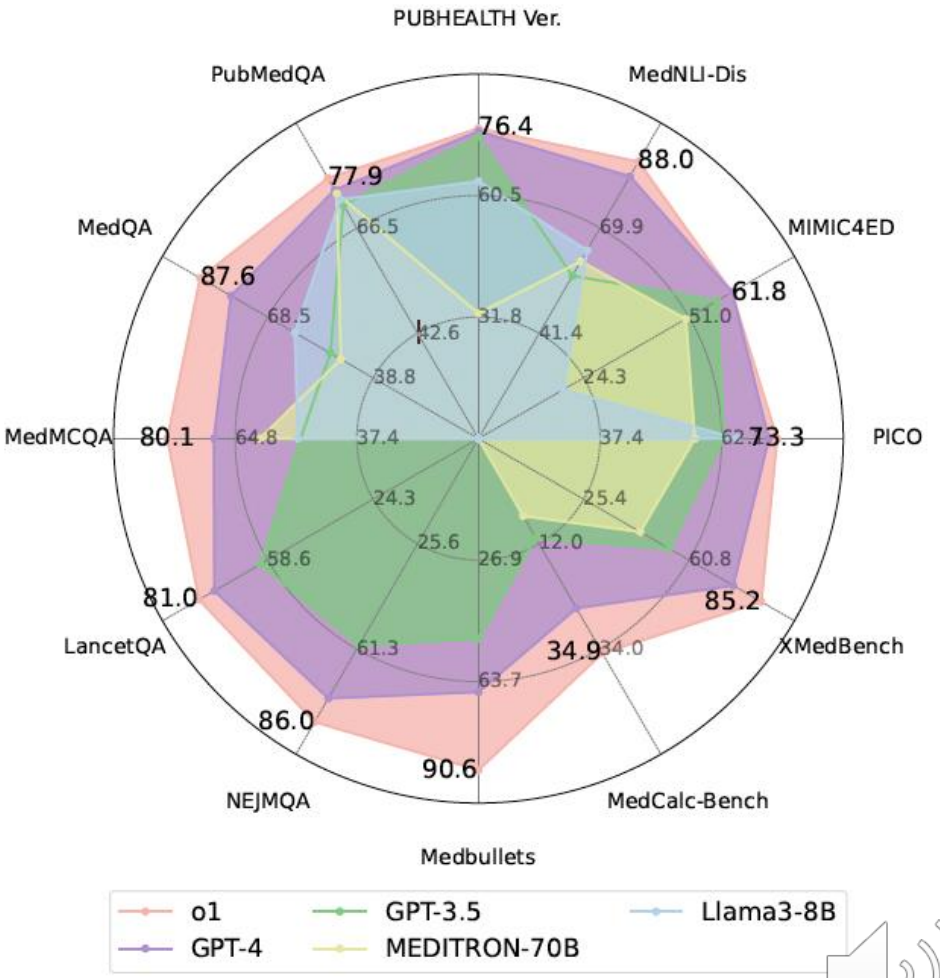
3.3 Generative Artificial Intelligence (2/3)

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

Answering clinical questions – clinicians

- Cancer
  - NCCN (Chen, 2023)
  - Guidelines (Ferber, 2024)
  - Questions (Rydzewski, 2024)
- Physician-generated questions (Goodman, 2023)
- OpenAI o1 outperformed most other LLMs on many medical questions (Xie, 2024)



Answering clinician questions (cont.)

- On multiple-choice genetics questions, scored comparable to humans (Duong, 2024)
- Almanac, using LLM framework augmented with retrieval capabilities from curated medical resources for medical guideline and treatment recommendations, showed significant improvement in performance compared with standard LLMs in factuality, completeness, user preference, and adversarial safety (Zakka, 2024)
- On questions seeking additional clinical evidence, ChatRWD using retrieval-augmented generation outperformed OpenEvidence and several general LLMs (Low, 2024)

OHSU

Answering clinical questions – patients

- 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)
- ChatGPT provided adequate information about radiation protection comparable to institutional websites (Jankowski, 2024)

Solving clinical cases

- NEJM clinicopathological cases (CPCs)
- Merck vignettes
- Mayo symptom checker
- Clinical vignettes

New England Journal of Medicine (NEJM) clinicopathologic conferences (CPCs)

- Long-standing use for evaluating AI, i.e., INTERNIST-1 (Miller, 1982)
- 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)
- 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 Google Med-PaLM 2 (McDuff, 2023)
  - Overall best DDx from LLM only, followed by generalist physicians with Med-PALM2, with search, and unassisted
- Recent analysis of newer cases found open-source Llama 3 able to score comparably to GPT-4 (Buckley, 2025)

Solving clinical cases (cont.)

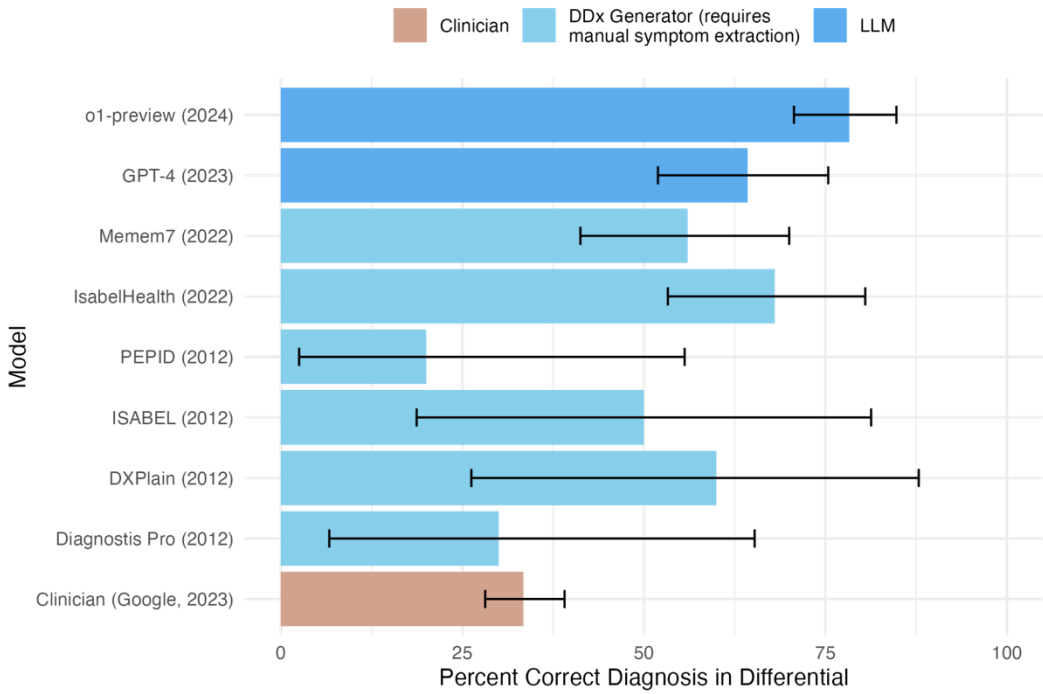
- For 194 diseases in Mayo Clinic Symptom Checker, ChatGPT-4 achieved 78.8% accuracy in making diagnosis, varying by clinical specialty (Chen, 2023)
- For 20 clinical cases, GPT-4 performed comparable to attending physicians and residents in diagnostic accuracy, correct clinical reasoning, and cannot-miss diagnosis inclusion (Cabral, 2024)
- Simulated cases created with Medical Information Mart for Intensive Care (MIMIC) records and used with open-source LLMs scored worse than physicians for 4 common abdominal conditions (Hager, 2024)
- For cases from PMC-Patients, adding lab results improved performance of formulating DDx (Bhasuron, 2025)

Clinical vignettes

- Diagnostic reasoning (Goh, 2024)
  - Based on 6 unpublished diagnostic cases developed to assess clinical decision support (Berner, 1994)
  - 50 physicians randomized to LLM or conventional information resources and assessed with diagnostic performance rubric
  - No statistical difference between physicians using LLM (76%) vs. conventional (74%) resources; LLM alone scored better than either (92%)
- Management reasoning (Goh, 2025)
  - Based on 5 cases adapted from *Grey Matters* podcast from American College of Physicians
  - 92 physicians randomized to LLM or conventional information resources and assessed with management performance rubric
  - Physicians using LLM scored better than physicians using conventional resources and no different from LLM alone

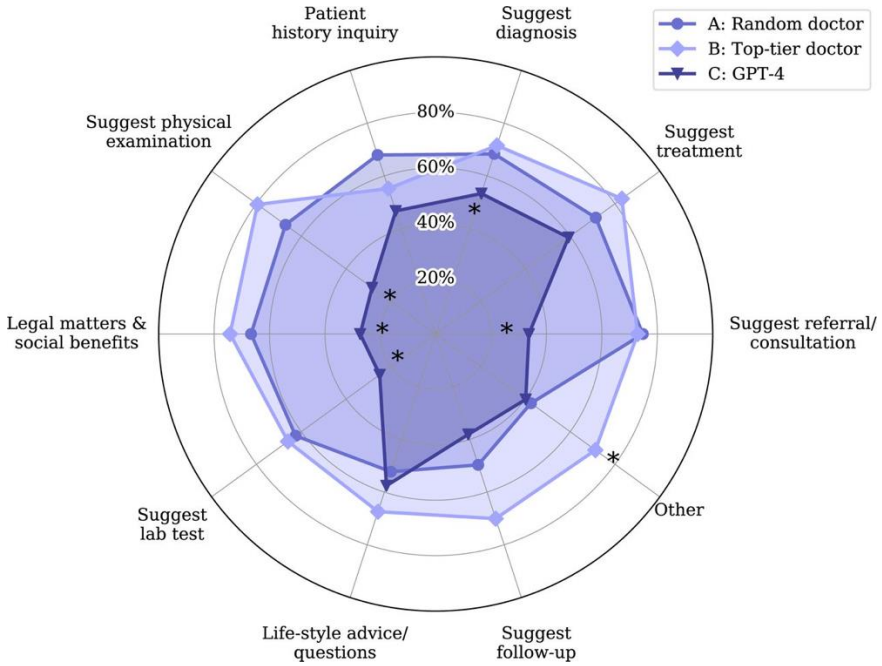
Clinical vignettes (cont.)

- “Superhuman performance” (?) with OpenAI o1-preview LLM scoring better than previous results with (Brodeur, 2024)
  - NEJM CPCs from (Kanjee, 2023)
  - Diagnostic probabilistic reasoning cases (Rodman, 2023)
  - NEJM Healer from (Cabral, 2024)
  - Landmark Diagnostic cases from (Goh, 2024)
  - Grey Matters management from (Goh, 2025)



Clinical vignettes (cont.)

- Comparison of responses from ChatGPT-4 and physicians for cases from Swedish family medicine specialist examination, scored by blinded reviewers (Arvidsson, 2024)
  - Higher scores for average physicians than GPT-4 or GPT-4o
- Conversational Reasoning Assessment Framework for Testing in Medicine (CRAFT-MD) focuses on natural dialogues, using simulated agents to interact with LLMs in controlled environment (Johri, 2025)
  - Performed worse in “conversational” than “examination-based” settings



Use of imaging models for imaging cases

- 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)
- For CXRs in emergency department (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)
- Imaging case vignettes with MCQs from *JAMA Clinical Challenges* and *NEJM Image Challenges*
  - GPT-4V without fine-tuning outperformed Gemini Pro, ChatGPT, and others (Han, 2024)
  - GPT-4V performed comparable to humans but can present flawed rationales (Jin, 2024)

Predictive tasks

- Cardiovascular disease (Han, 2024)
- ED acuity (Williams, 2024) and predicting admissions (Glicksberg, 2024)
- Rare disease diagnosis (Zelin, 2024)
- Designing and validating novel antibiotics (Swanson, 2024)
- Diagnose specific infections, autoimmune disorders, vaccine responses, and disease severity differences based on T and B cell receptor sequences (Zaslavsky, 2025)

Summarization – EHR data for clinicians

- Extracting details from discharge summaries with 81% accuracy (Ellershaw, 2024)
- For radiology reports, patient questions, progress notes, and doctor-patient dialogue, LLM summaries found preferable to human summaries (Van Veen, 2024)
- LLM-generated emergency medicine (EM)-to-inpatient physician (IP) handoff notes determined superior compared with physician-written summaries but marginally inferior in usefulness and safety (Hartman, 2024)
- LLM answering questions from clinical notes in 3 languages in high agreement with humans (Menezes, 2025)

OHSU

Summarization – scientific papers

- 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)
- PaperQA2 summarized topics comparable to Wikipedia and identified contradictions in papers (Skarlinski, 2024)

OHSU

Summarization – for patients

- 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)
- Lay language summaries for research studies found more accessible and transparent (Shyr, 2024)
- Transforming clinical notes for patients rated patient-friendly but 44% not entirely complete and 18% found safety concerns for incomplete or inaccurate information (Zaretsky, 2024)
- Generating patient-friendly summaries of radiology reports understandable to patients (Park, 2024)

Drafting replies to patients

- ChatGPT-3.5 wrote patient clinic letters with high level of correctness and measure of “humanness” (Ali, 2023)
- Fine-tuned 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)
- For drafting replies, LLM not associated with reduced time on writing a reply but was associated with longer read time, longer replies, and perceived value in making a more compassionate reply Tai-Seale, 2024)
- Help patients draft messages to health system (Liu, 2024)

Drafting replies (cont.)

- GPT-4-generated patient portal message responses achieve comparable levels of empathy, relevance, and readability to those found in typical responses according to providers (Kaur, 2024)
- Patient Advice Message Chatbot drafted reply to incoming messages from patient portal in 9 clinics for nurses, medical assistants (MAs), and clinicians (physicians and advance practice clinicians [APCs]) (English, 2024)
- Satisfaction with AI-generated responses to medical questions in EHR (Kim, 2024; Cavalier, 2025) but reduced when patient notified response was AI-generated (Cavalier, 2025)

Patient documentation-related tasks

- De-identification
  - Radiology reports (Chambon, 2023)
  - Discharge summaries (Altalla, 2025)
- Identifying social determinants of health (Guevara, 2024)
- Extraction of Severe Sepsis and Septic Shock Management Bundle (SEP-1) quality measure (Boussina, 2024)
- But poor performance in some tasks
  - ICD-10-CM and CPT-4 coding (Soroush, 2024)
  - Using different drug names (generic vs. trade) leads to differing performance on standard datasets (Gallifant, 2024)

OHSU

Other tasks

- Discover math errors in scientific papers (Mollick, 2024; Brean, 2024)
- Psychotherapy (Hatch, 2025)
- Pediatric medication dosage errors (Levin, 2025)
- Systematic reviews – screening for papers (Cao, 2025) and extracting data from them (Khan, 2025)

Work productivity

- 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
- Predicting company earnings (Kim, 2024; Shaffer, 2024)
- Software engineering – improved productivity (Cui, 2024)
- In materials science, AI-assisted researchers discover 44% more materials, resulting in a 39% increase in patent filings and a 17% rise in downstream product innovation (Toner-Rodgers, 2024)

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
- GPT-4 performed worse than humans in abstraction and analogy (Moskvichev, 2023) and on abstraction and reasoning corpus (ARC; Chollet, 2019) (Mitchell, 2023) but newer o3 has “solved” problem (Mitchell, 2024)