3.2 Generative Artificial Intelligence (1/3)

What is Biomedical and Health Informatics? - http://informatics.health/
William Hersh, MD, FACMI, FAMIA, FIAHSI
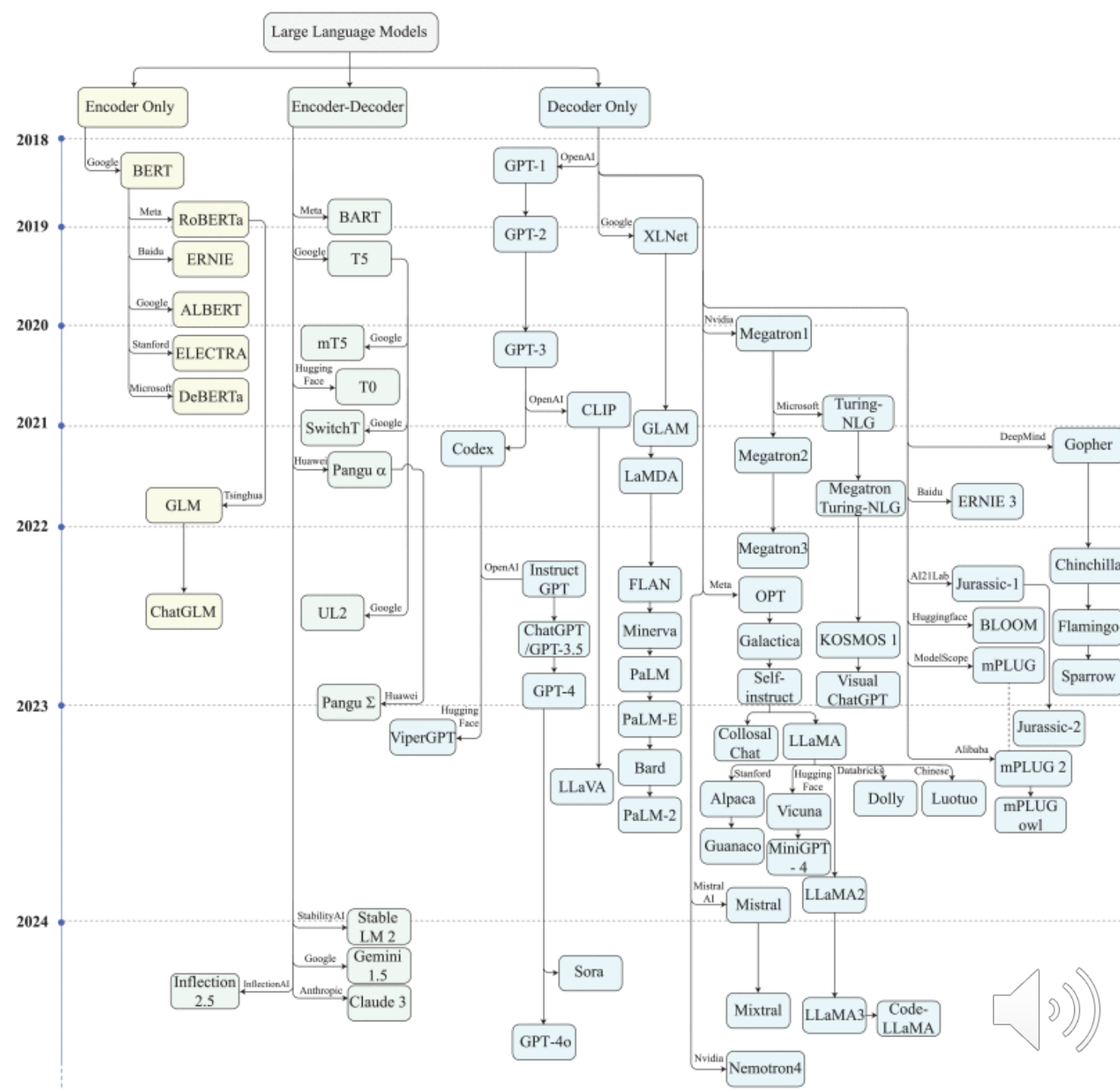Copyright 2025

Generative AI

- Overview
- Systems and usage
- Prompting
- Research results
- Downsides
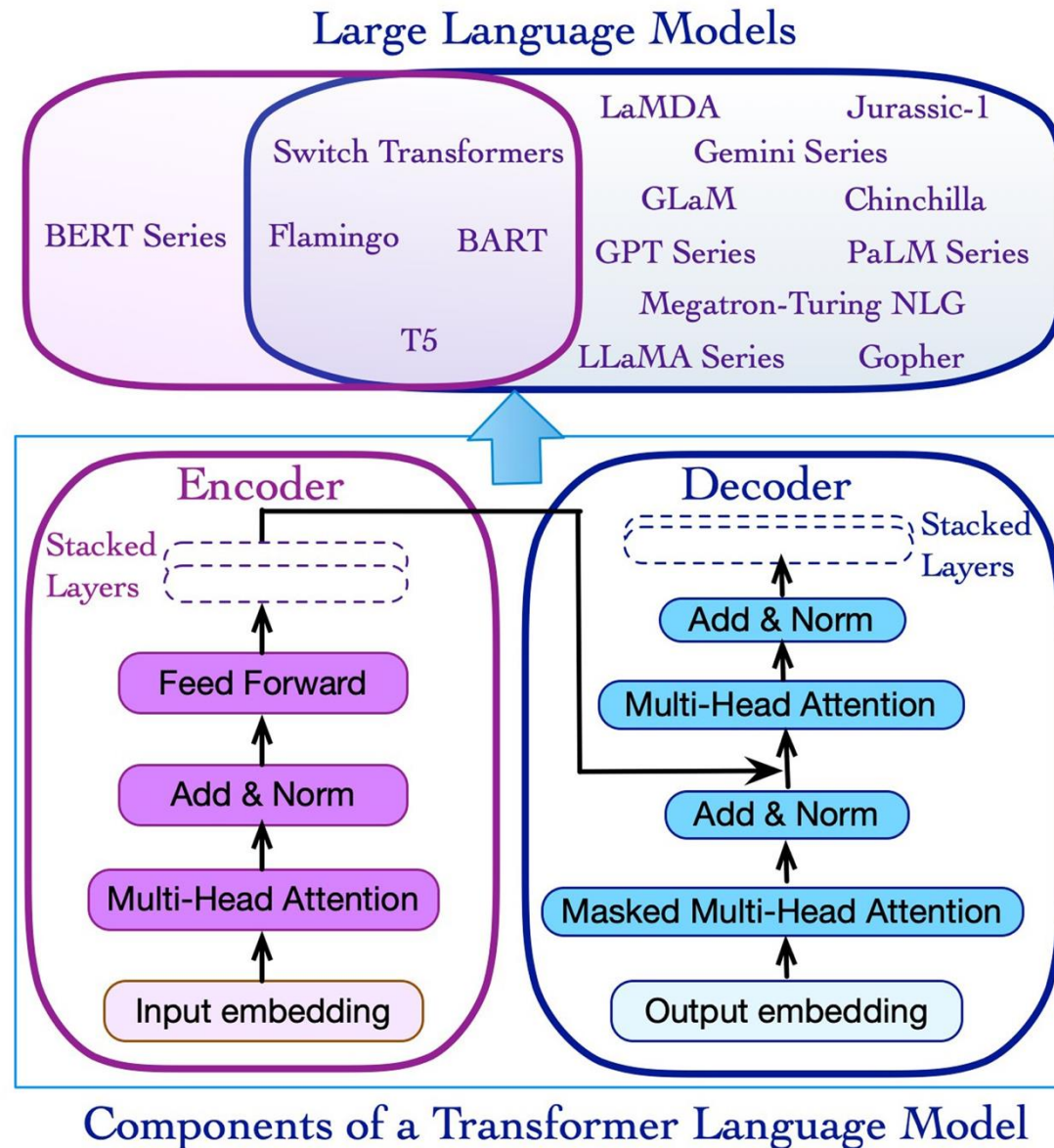- Impact of education

Overview

- Thrust into public eye with November 30, 2022 release of ChatGPT by OpenAI
  - Based on large language models (LLMs) with number of new advances
- Impacts in all aspects of society, not only healthcare and education but also news, politics, and more
  - Two-year retrospective (Lenharo, 2024)
- Early experience exemplified by 2023 Dictionary.com word of the year: *hallucinate* (Norlen, 2023)
  - Problem improving but persists (Jones, 2025)
- Inner workings unknown to humans (Wolfram, 2023)

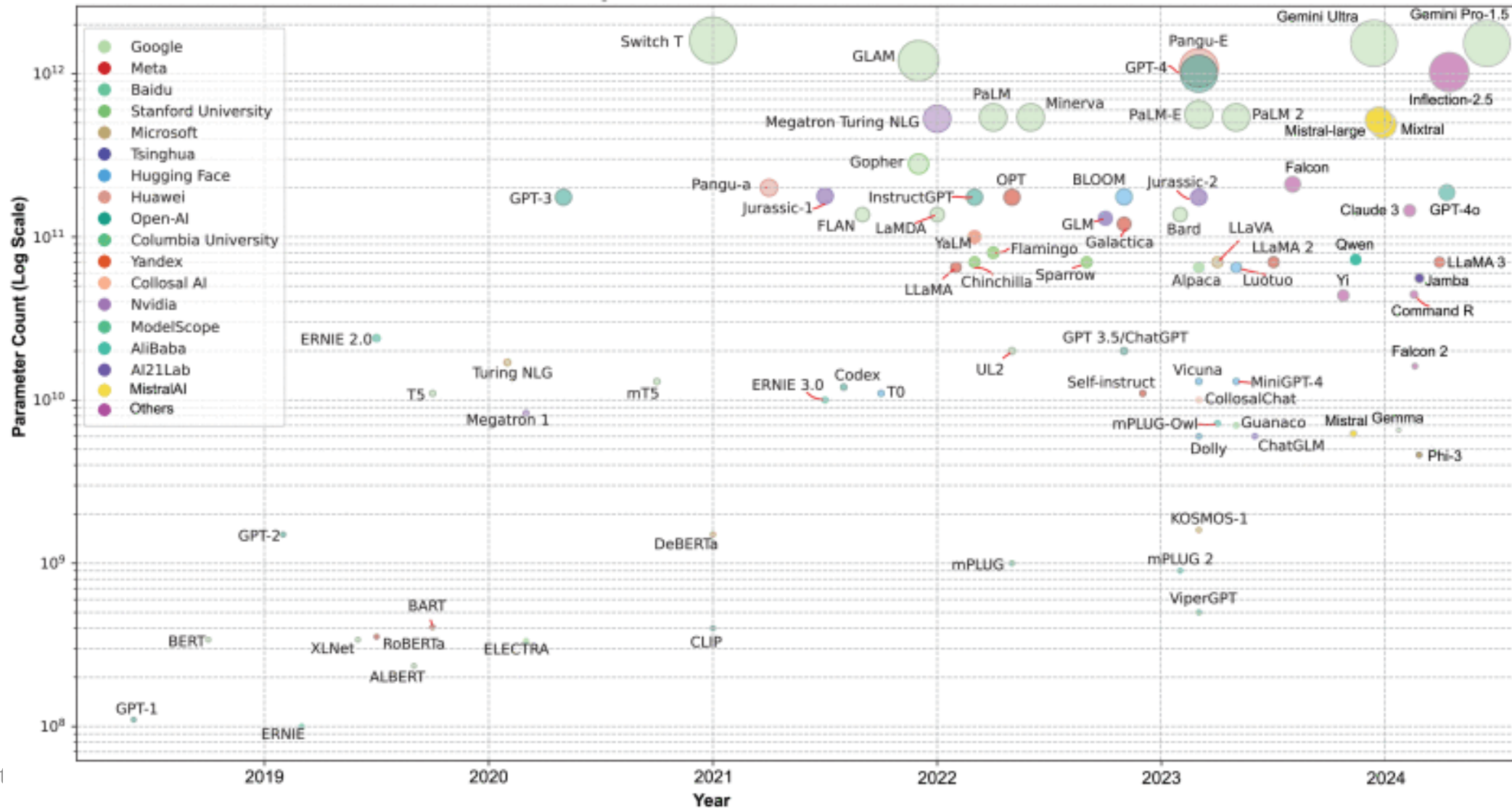# Lineage of current LLMs (Shao, 2024)

- Major commercial LLM families
  - GPT (OpenAI)
  - Gemini (Google)
  - Claude (Anthropic)
  - LLaMA (Meta)
  - Mistral
  - And more...
- All are decoder transformers that process prompts to decode from model to generate text

LLMs are decoder transformers (Sahoo, 2024)

Large Language Models

Switch Transformers
BERT Series   Flamingo   BART
T5

LaMDA   Jurassic-1
Gemini Series
GLaM   Chinchilla
GPT Series   PaLM Series
Megatron-Turing NLG
LLaMA Series   Gopher

Encoder
Stacked Layers
Feed Forward
Add & Norm
Multi-Head Attention
Input embedding

Decoder
Stacked Layers
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Output embedding

Components of a Transformer Language Model

Increasing sizes of models (Shao, 2024)

Advances beyond basic LLMs

- Chain of thought (Wei, 2022) – trained to reason in steps rather than all at once, e.g.,
  - OpenAI o1 and beyond (Xie, 2024)
  - DeepSeek (Gibney, 2025)
- Agentic AI – autonomous agents designed to perform specific tasks step by step instead of all at once (Purdy, 2024)
  - Different functions in an EHR (Zou, 2025)

- Foundational models – fine-tuned to specific data types or domains, e.g.,
  - Genome sequences (Nguyen, 2024; Callaway, 2025)
  - Ophthalmology (Qiu, 2024)
  - Dermatology (Gui, 2024)
  - DNA (Benegas, 2025)
  - Imaging (Zhang, 2025; Zhao, 2025)
  - Precision oncology (Xiang, 2025)
  - Laryngology (Setzen, 2025)
  - Medical literature mining (Wang, 2025)
  - Disease diagnosis (Liu, 2025)

Usage

- Survey of college students found 56% stated use of AI on assignments or exams, with 54% agreeing that use on college coursework counts could constitute cheating or plagiarism (Nam 2023)
- Survey of teachers, students, and parents found 50% stated use of generative AI frequently, 18-33% occasionally, and less than quarter had never used (Impact Research, 2024)
- Survey of US teens found 26% used for homework (Sidoti, 2025)
- Survey of working adults (Bick et al. 2024)
  - 39% overall use
  - 24% used in work at least once a week
  - About 10% used every day at work

Usage (cont.)

- Kaiser Family Foundation (KFF) Health Misinformation Tracking Poll (Presiado, 2024) – of 2428 US adults varying by age, race/ethnicity, and geography reporting use or interaction with AI
  - Several times a day – 11%
  - About once a day – 8%
  - Several times a week – 14%
  - Less often – 30%
  - Never – 35%
- Survey of over 1000 physicians from UK (Blease et al. 2024)
  - 20% reported using generative AI tools in clinical practice
  - Among those who used, 29% reported used to generate documentation after patient appointments and 28% to develop differential diagnosis

AMA survey of physicians conducted in November 2024 (AMA, 2025)

- 68% of physicians surveyed see definite or some advantage to using AI tools, up from 65% in 2023
- 66% indicated currently use AI in practice, up significantly from 38% in 2023
- Top area of opportunity, according to 57% of physicians surveyed, was "addressing administrative burden through automation," up from 56% in 2023
- Top attributes required to advance physician adoption of AI tools were
  - Designated feedback channel - 88%
  - Data privacy assurances - 87%
  - EHR integration - 84%

Prompting

- Two main approaches
  - Web interface, e.g., ChatGPT
  - Application programming interface (API)
- Much advice and education available – books, articles, Web sites, courses, etc. – some general ones
  - OpenAI
  - AI Cheat Sheets
  - For students (Mollick, 2023) and instructors (Mollick, 2024)
- May be a "poor user interface" (Morris, 2024)

Prompting for medical professionals (Meskó, 2023)

SPECIFIC PROMPT RECOMMENDATIONS

1. The more specific your prompt is, the more accurate the response is likely to be.

Example
Less specific:
"Tell me about heart disease."
More specific:
"What are the most common risk factors for coronary artery disease?"

2. Describe your setting and provide the context!

Example
I'm writing an article about tips and tricks for ChatGPT prompt engineering for people working in healthcare. Can you please list a few of those tips and tricks with some specific prompt examples?"

3. Experiment with different prompt styles!

Example
Direct Question:
"What are the symptoms of COVID-19?"
Request for List:
"List all the potential symptoms of COVID-19."
Request for Summary:
"Summarize the key symptoms and progression of COVID-19."
"Explain the symptoms of COVID-19 like I'm five."

4. Identify the overall goal of your prompt first!

Example
"I'd like to get a short list of 5 ideas for a Youtube video on the future of healthcare AI."

5. Ask ChatGPT to play roles!

Example
"Act as a Data Scientist and explain Prompt Engineering to a physician."
"Act as my nutritionist and give me tips about a balanced Mediterranean diet."

6. Iterate and refine your question and ask ChatGPT to modify the output based on its previous response.

Example
Initial prompt:
"How are you feeling today?"
Refined prompt:
"On a scale of 1-10, how would you rate your stress levels today and what specific event contributed to it?"

7. Use threads

You can easily go back to a specific discussion by clicking on the proper thread in the left column. This way, you don't have to start all over again but can just continue a discussion you already had with ChatGPT.

8. Ask Open-ended Questions as those often yield more comprehensive responses.

Example
Closed question:
"Is exercise important for patients with osteoporosis?"
Open question:
"How does regular physical activity benefit patients with osteoporosis?"

9. Request specific examples

If there is something you are not satisfied with or don't understand based on its response, first of all, tell ChatGPT that you don't understand the answer and ask it to provide an example.

Initial prompt:
"Could you explain the common side effects of this medication?"
Refined prompt:
"I'm not clear about the side effects of this medication. Can you provide specific examples of common side effects patients have experienced?"

10. If you're asking about a process or timeline, specify that in your prompt.

Example
Without time reference:
"Describe the healing process after knee surgery."
With time reference:
"What can a patient typically expect during the first six weeks of healing after knee surgery?"

+1 And the +1, Set Realistic Expectations:

Example
Unrealistic Prompt:
"What's the latest research published this month about Alzheimer's?"
Realistic Prompt:
"What were some of the major research breakthroughs in Alzheimer's treatment up until 2021?"
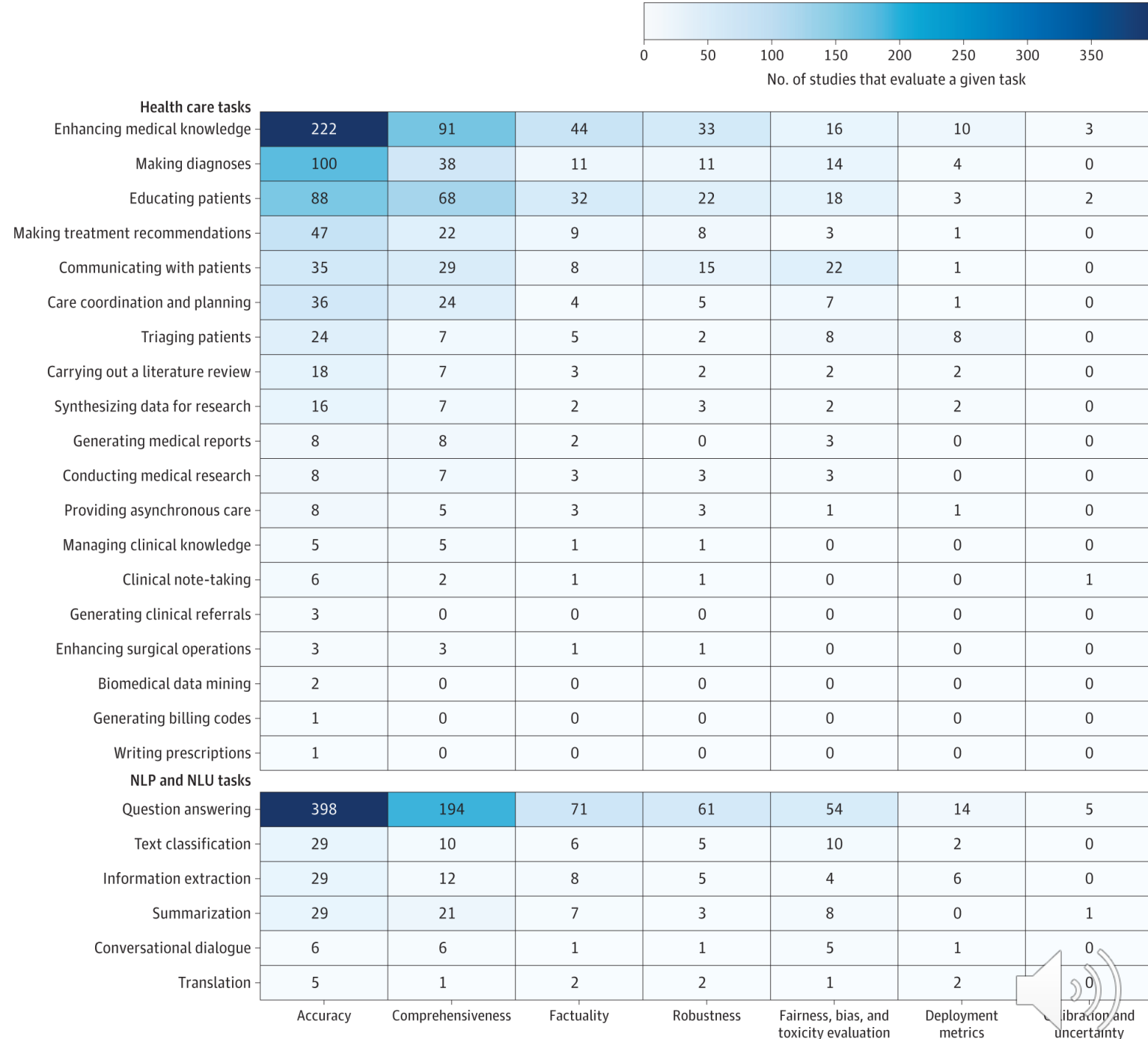
WhatIs3.2

Implementing LLMs in healthcare settings

- Security essential to prevent leakage of protected health information (PHI) into models (Ng, 2025)
- Governance deemed critical at research-oriented academic medical centers (Idnay, 2025)
- Some case studies
  - Stanford University (Langlotz, 2024)
  - New York University-Langone (Malhorta, 2025)

Gen AI applications and research results

- Large number of studies, some of which are preprints, i.e., not peer-reviewed

- Most studies test models with benchmark datasets, i.e., good for comparing models but losing value over time and no indication of real-world efficacy

- Most systems, especially large commercial ones, are "black boxes," i.e., no way to "see" inside and constantly changing

- Systematic reviews show large and ever-growing volume of studies, making such reviews out of date quickly

Systematic review of healthcare applications of LLMs (Bedi, 2024)

No. of studies that evaluate a given task

Health care tasks

	Accuracy	Comprehensiveness	Factuality	Robustness	Fairness, bias, and toxicity evaluation	Deployment metrics	Calibration and uncertainty
Enhancing medical knowledge	222	91	44	33	16	10	3
Making diagnoses	100	38	11	11	14	4	0
Educating patients	88	68	32	22	18	3	2
Making treatment recommendations	47	22	9	8	3	1	0
Communicating with patients	35	29	8	15	22	1	0
Care coordination and planning	36	24	4	5	7	1	0
Triaging patients	24	7	5	2	8	8	0
Carrying out a literature review	18	7	3	2	2	2	0
Synthesizing data for research	16	7	2	3	2	2	0
Generating medical reports	8	8	2	0	3	0	0
Conducting medical research	8	7	3	3	3	0	0
Providing asynchronous care	8	5	3	3	1	1	0
Managing clinical knowledge	5	5	1	1	0	0	0
Clinical note-taking	6	2	1	1	0	0	1
Generating clinical referrals	3	0	0	0	0	0	0
Enhancing surgical operations	3	3	1	1	0	0	0
Biomedical data mining	2	0	0	0	0	0	0
Generating billing codes	1	0	0	0	0	0	0
Writing prescriptions	1	0	0	0	0	0	0

NLP and NLU tasks

	Accuracy	Comprehensiveness	Factuality	Robustness	Fairness, bias, and toxicity evaluation	Deployment metrics	Calibration and uncertainty
Question answering	398	194	71	61	54	14	5
Text classification	29	10	6	5	10	2	0
Information extraction	29	12	8	5	4	6	0
Summarization	29	21	7	3	8	0	1
Conversational dialogue	6	6	1	1	5	1	0
Translation	5	1	2	2	1	2	0

Dimension of evaluation

Categories and tasks for generative AI in healthcare (Shah, 2025)

WhatIs3.2
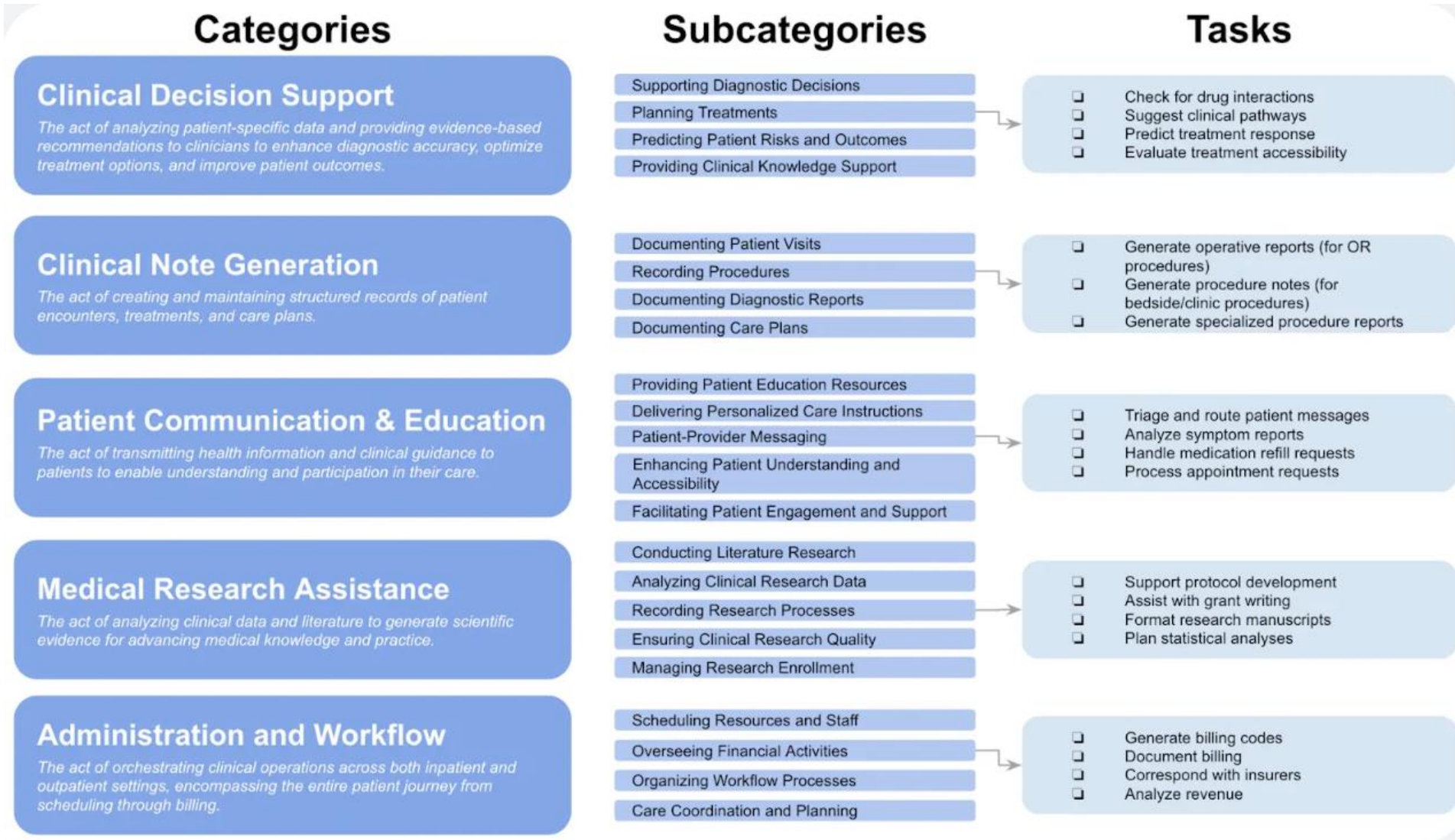
Specific research results

- Passing medical board exams
- Answering questions – clinicians and patients
- Solving clinical cases
- Prediction
- Summarization
- Generating text for patients
- Aiding patient documentation
- Educational assessment
- Beyond healthcare
  - Work productivity
  - Educational use

Medical board exams

- US Medical Licensing Exam (USMLE)
- Post-residency exams in Israel
- Other areas
  - Clinical informatics
  - Radiology
  - Family medicine
  - Other specialties

USMLE "arms race"

- MedQA dataset based on sample multiple-choice questions (12,273 in English, more in Chinese) from three-part medical exam required for licensing in US (Jin, 2021)

Question	A 27-year-old male presents to urgent care complaining of pain with urination. He reports that the pain started 3 days ago. He has never experienced these symptoms before. He *denies gross hematuria or pelvic pain*. He is sexually active with his girlfriend, and they consistently use condoms. When asked about recent travel, he admits to recently returning from a "boys' trip" in Cancun where he had *unprotected sex* 1 night with a girl he met at a bar. The patient's medical history includes type I diabetes that is controlled with an insulin pump. His mother has rheumatoid arthritis. The patient's temperature is 99 °F(37.2 °C), blood pressure is 112/74 mmHg, and pulse is 81/min. On physical examination, there are no lesions of the penis or other body rashes. No costovertebral tenderness is appreciated. A urinalysis reveals no blood, glucose, ketones, or proteins but is *positive for leukocyte esterase*. A urine microscopic evaluation shows a *moderate number of white blood cells* but no casts or crystals. A urine culture is negative. Which of the following is the most likely cause for the patient's symptoms?
Options	**A: Chlamydia trachomatis**, B: Systemic lupus erythematosus, C: Mycobacterium tuberculosis, D: Treponema pallidum
Evidence	At least one-third of male patients with *C. trachomatis* urethral infection have *no evident signs or symptoms of urethritis*. … Such patients generally have *pyuria* …, a *positive leukocyte esterase test*, …
Question	A 57-year-old man presents to his primary care physician with a 2-month history of *right upper and lower extremity weakness*. He noticed the weakness when he started falling far more frequently while running errands. Since then, he has had *increasing difficulty* with walking and lifting objects. His past medical history is significant only for well-controlled hypertension, but he says that some members of their *family have had musculoskeletal problems*. His right upper extremity shows *forearm atrophy* and *depressed reflexes* while their right lower extremity is *hypertonic with a positive Babinski sign*. Which of the following is most likely associated with the cause of this patient's symptoms?
Options	A: HLA-B8 haplotype, B: HLA-DR2 haplotype, **C: Mutation in SOD1**, D: Mutation in SMN1, E: Viral infection
Evidence	1. The manifestations of ALS … *insidiously* developing *asymmetric weakness*, usually first evident distally in one of the limbs. 2. … *hyperactivity of the muscle-stretch reflexes (tendon jerks)* and, often, *spastic resistance to passive movements* … 3. *Familial ALS (FALS)*… clinically indistinguishable from sporadic ALS… Genetic studies have identified mutations in multiple genes, including cytosolic enzyme *SOD1*…

What

OHSU

USMLE "arms race"

- First "passing" score (>60%) obtained with GPT-3.5 (Kung, 2023)
  - Even on "soft skills" (Brin, 2023)
- Subsequent improvement in LLMs and prompting have led to >90% scores
  - GPT-4 with specialized prompting (Nori, 2023)
  - Gemini (Corrado, 2024; Saab, 2024)
  - OpenAI o1-preview with simple prompting (Nori, 2024; Horvitz, 2024)
- Time to end use of this benchmark (Raji, 2025)?

MedQA Test Accuracy Over Time
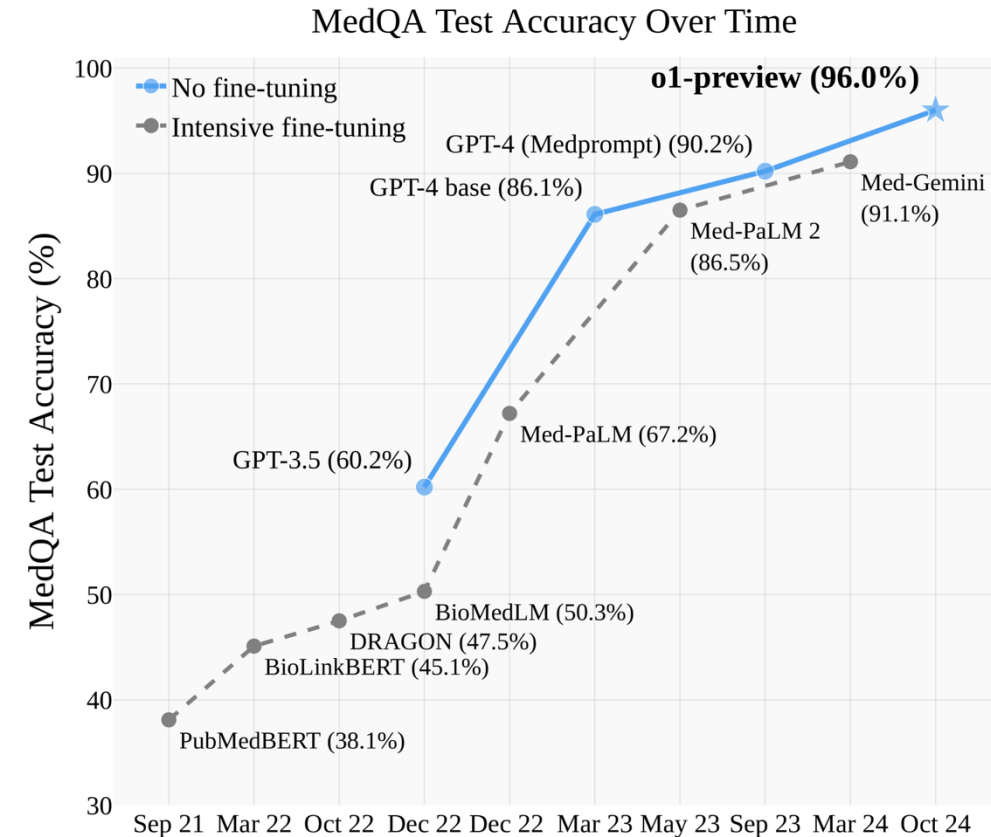
Post-residency board exams in Israel (Katz, 2024)
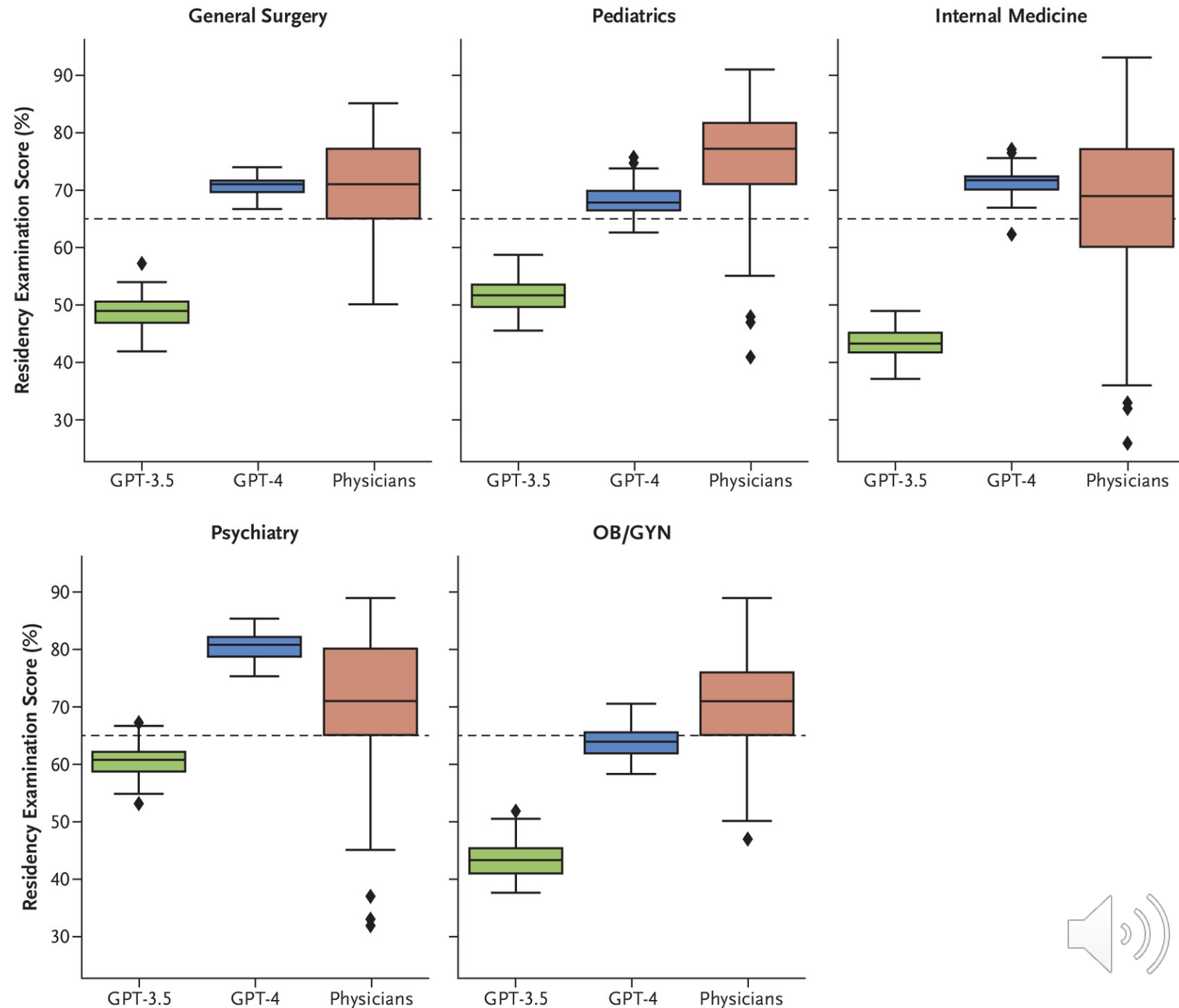
Other board and in-training exams

- Clinical informatics (Kumah-Crystal, 2023)
- Radiology
  - Board exams for GPT-4 (Bhayana, 2023; Bhayana, 2023) and Llama 3 (Adams, 2024)
  - In-training exam (Payne, 2024)
- Family medicine in-training exams (Hanna, 2024; Wang, 2024)
- Subspecialty self-assessments
  - Ophthalmology (Teebagey, 2023)
  - Nephrology (Wu, 2024)