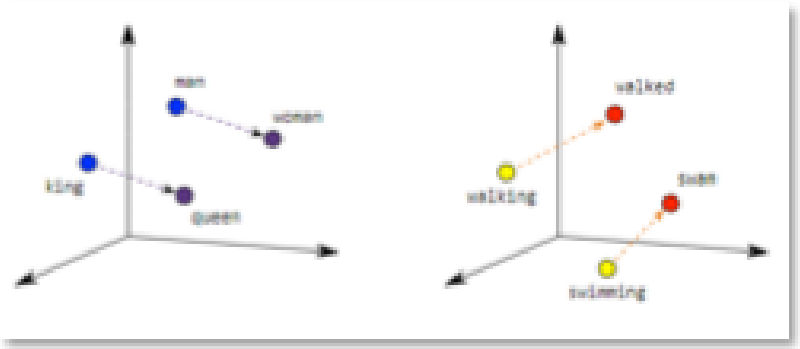
Data Science and AI for Medicine Training School

TRAINING: LLMs in Medicine

SPEAKER: Sanddhya Jayabalan

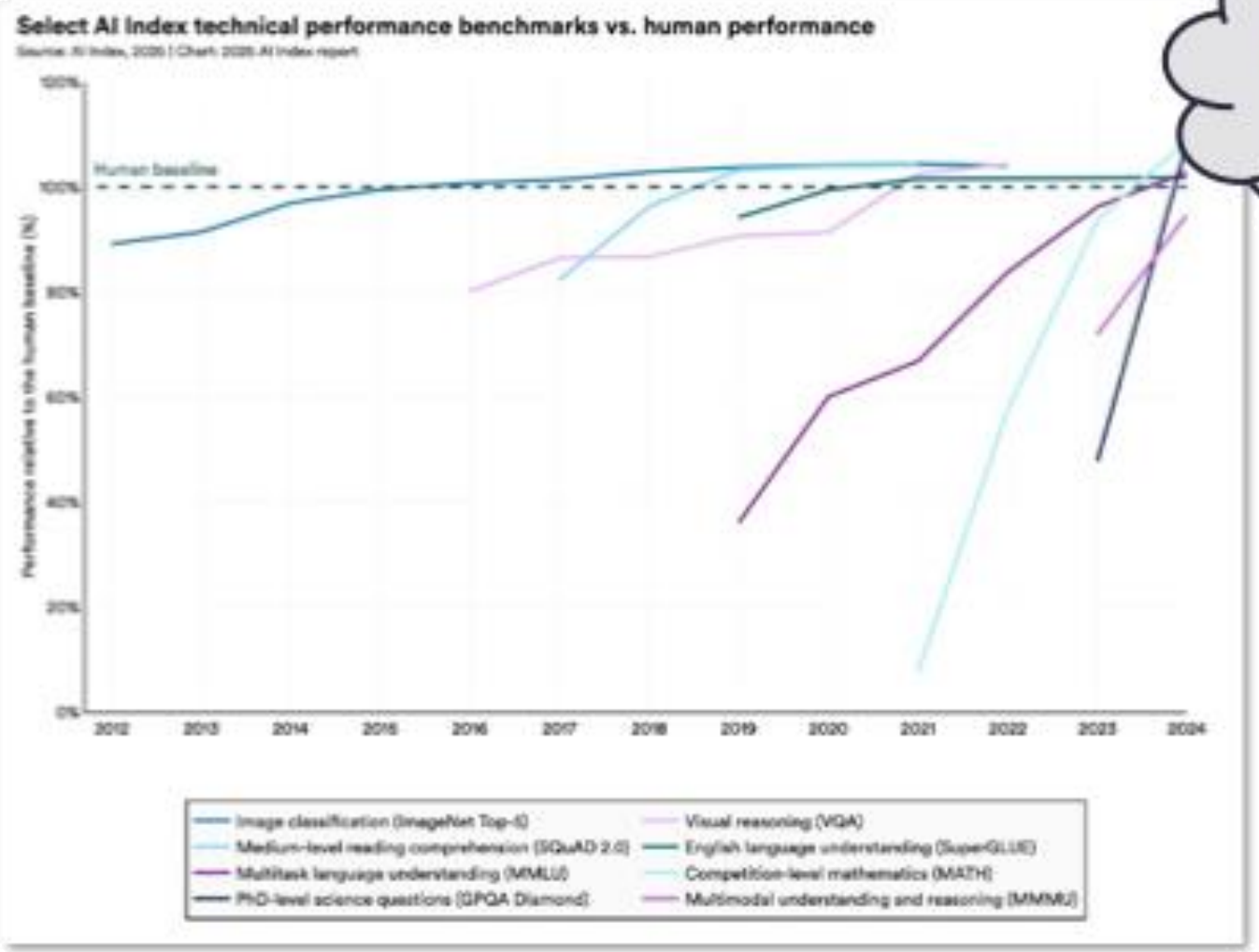
GEFÖRDERT VOM
Bundesministerium für Forschung, Technologie und Raumfahrt

SACHSEN
Diese Maßnahme wird gefördert durch die Bundesregierung aufgrund eines Beschlusses des Deutschen Bundestages. Diese Maßnahme wird mitfinanziert durch Steuermittel auf der Grundlage des von den Abgeordneten des Sächsischen Landtags beschlossenen Haushaltes.



Large Language Models.

LLMs are highly performant – zero-shot!



https://hai.stanford.edu/ai-index/2025-ai-index-report

Source: https://www.tum.de/en/news-and-events/all-news/press-releases/details/chatgpt-gruender-sam-altmann-an-der-tum

Medicine.

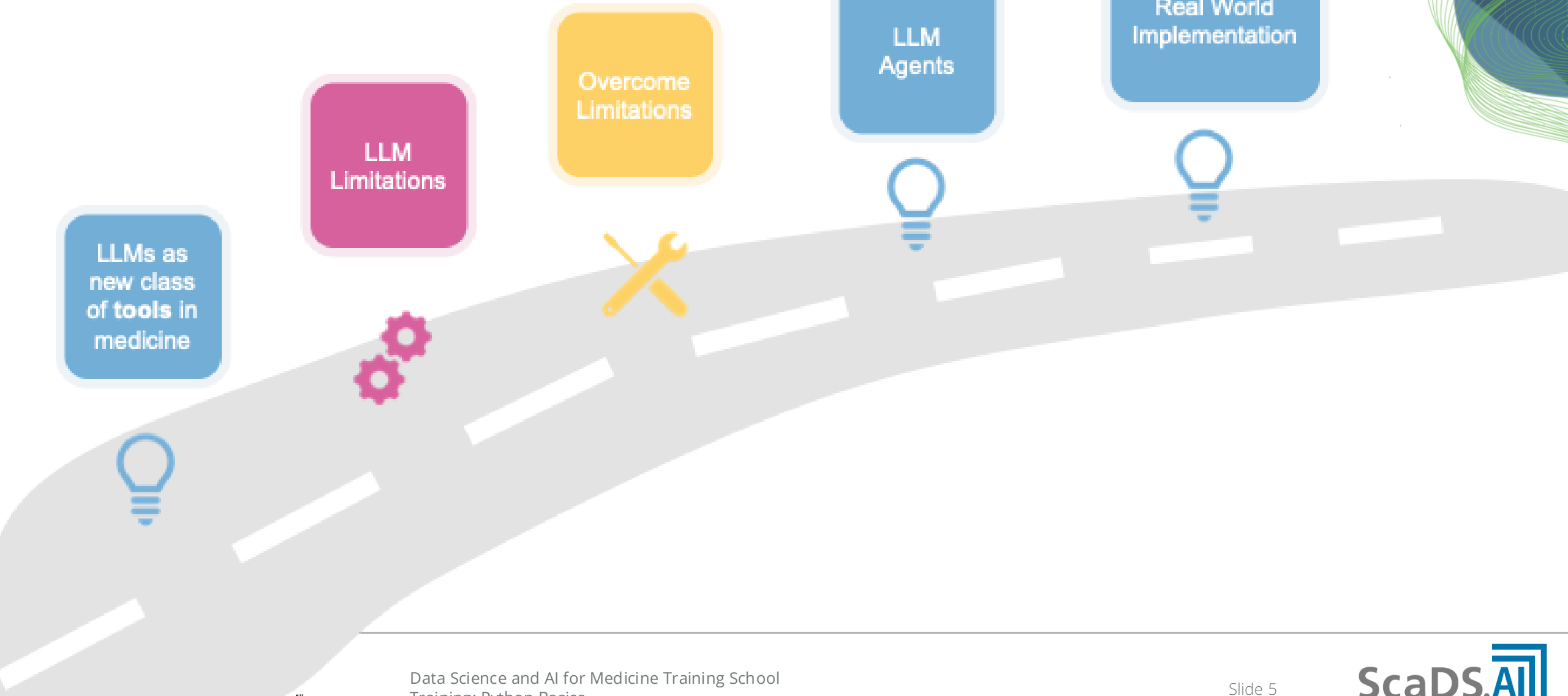
80% of medical data is unstructured

Growing knowledge

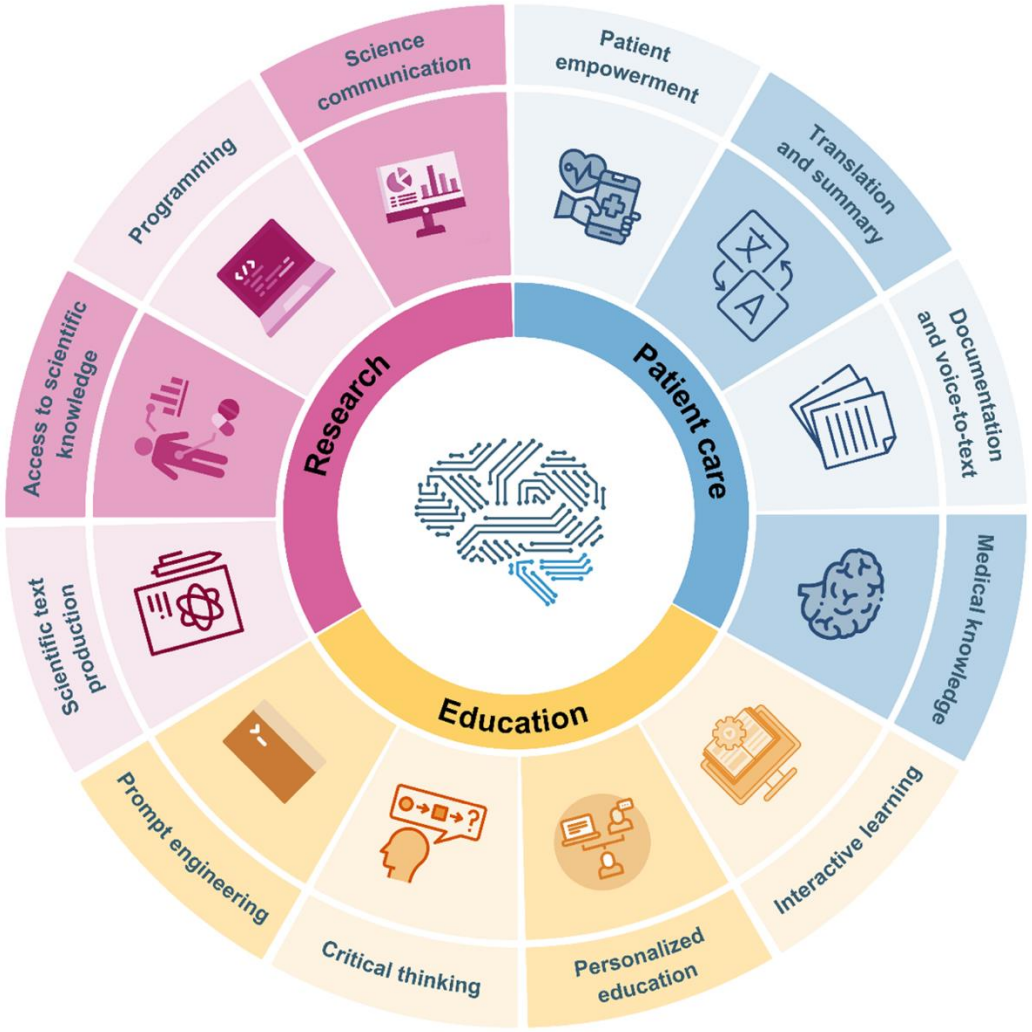
Narrative, free text contains valuable insights

Need for personalized medicine

Today's Roadmap



Large Language Models: A new class of tools in medicine



Clusmann, J., Kolbinger, F.R., Muti, H.S. *et al.* The future landscape of large language models in medicine. *Commun Med* **3**, 141 (2023). https://doi.org/10.1038/s43856-023-00370-1

Large Language Models: A new class of tools in medicine

LLMs structure unstructured data

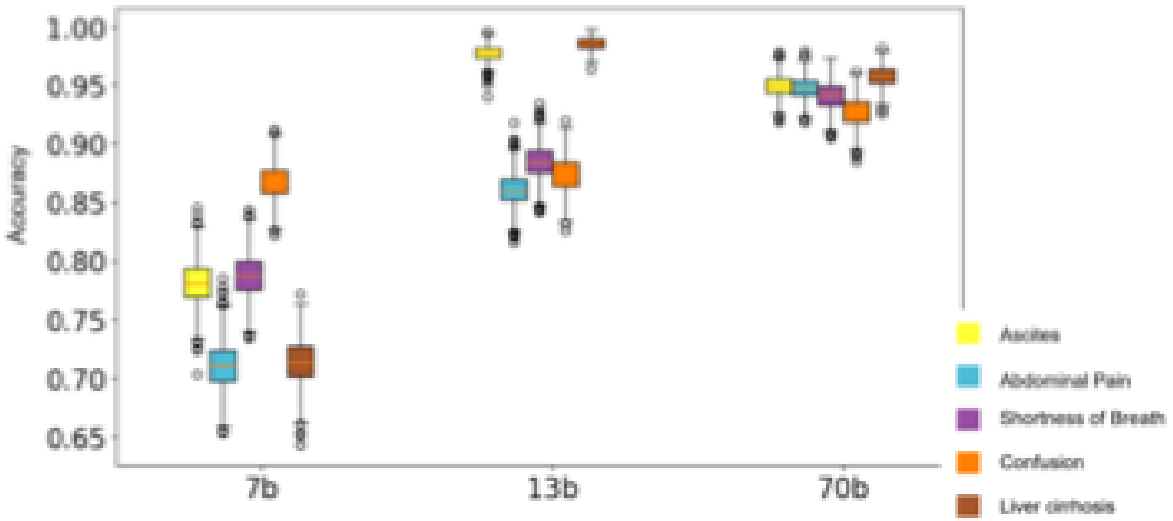
- Cancer registries
- Electronic Patient Records
- Medical Coding
- …

LLM-AIx

npj | digital medicine

Privacy-preserving large language models for structured medical information retrieval

Isabella Catharina Wiest, Dyke Ferber, Jiefu Zhu, Marko van Treeck, Sonja K. Meyer, Radhika Juglan, Zunamys I. Carrero, Daniel Paech, Jens Kleesiek, Matthias P. Ebert, Daniel Truhn & Jakob Nikolas Kather



Large Language Models: A new class of tools in medicine

LLMs summarize medical text

- Chart review
- Actionable guideline summaries
- …

Doctor LLM is ready to see you now.

https://greator.com/beziehungsprobleme/

NEJM AI

NEJM AI 2025;2(9)
DOI: 10.1056/AIoa2400802

ORIGINAL ARTICLE

Randomized Trial of a Generative AI Chatbot for Mental Health Treatment

Michael V. Heinz, M.D.,[1,2] Daniel M. Mackin, Ph.D.,[1,2] Brianna M. Trudeau, B.A.,[1] Sukanya Bhattacharya, B.A.,[1] Yinzhou Wang, M.S.,[1] Haley A. Banta,[1] Abi D. Jewett, B.A.,[1] Abigail J. Salzhauer, B.A.,[1] Tess Z. Griffin, Ph.D.,[1] and Nicholas C. Jacobson, Ph.D.[1,3,4]

Received: August 11, 2024; Revised: November 18, 2024; Accepted: February 2, 2025; Published March 27, 2025

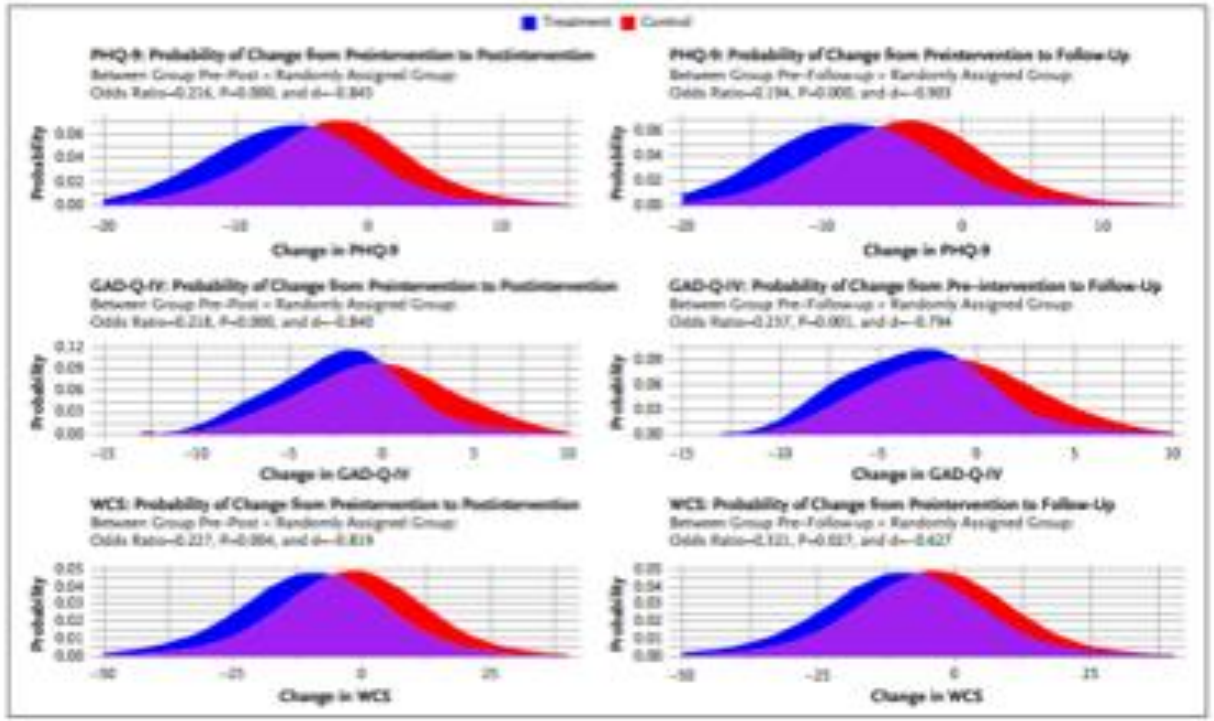
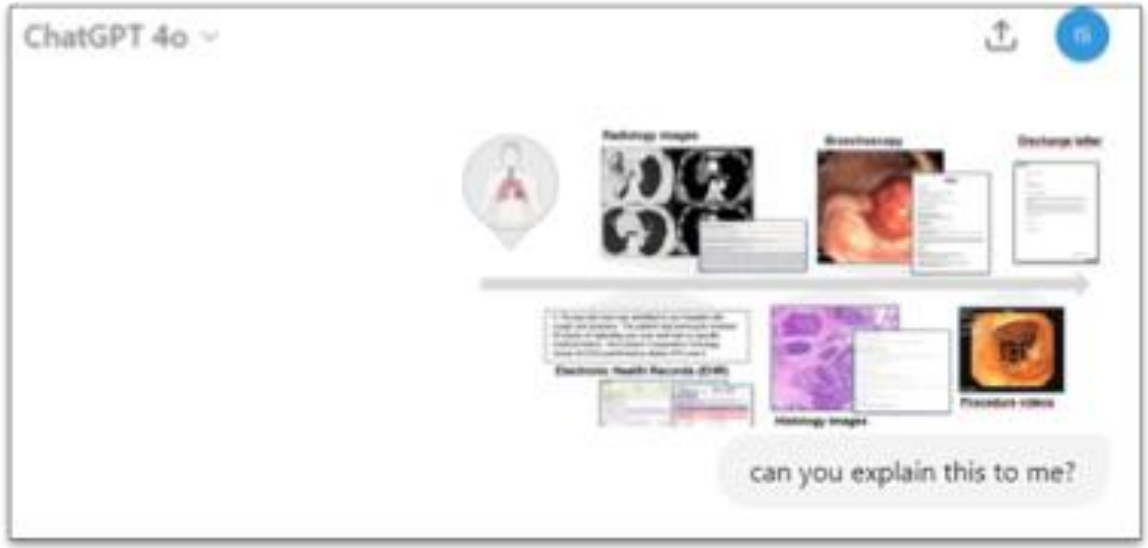
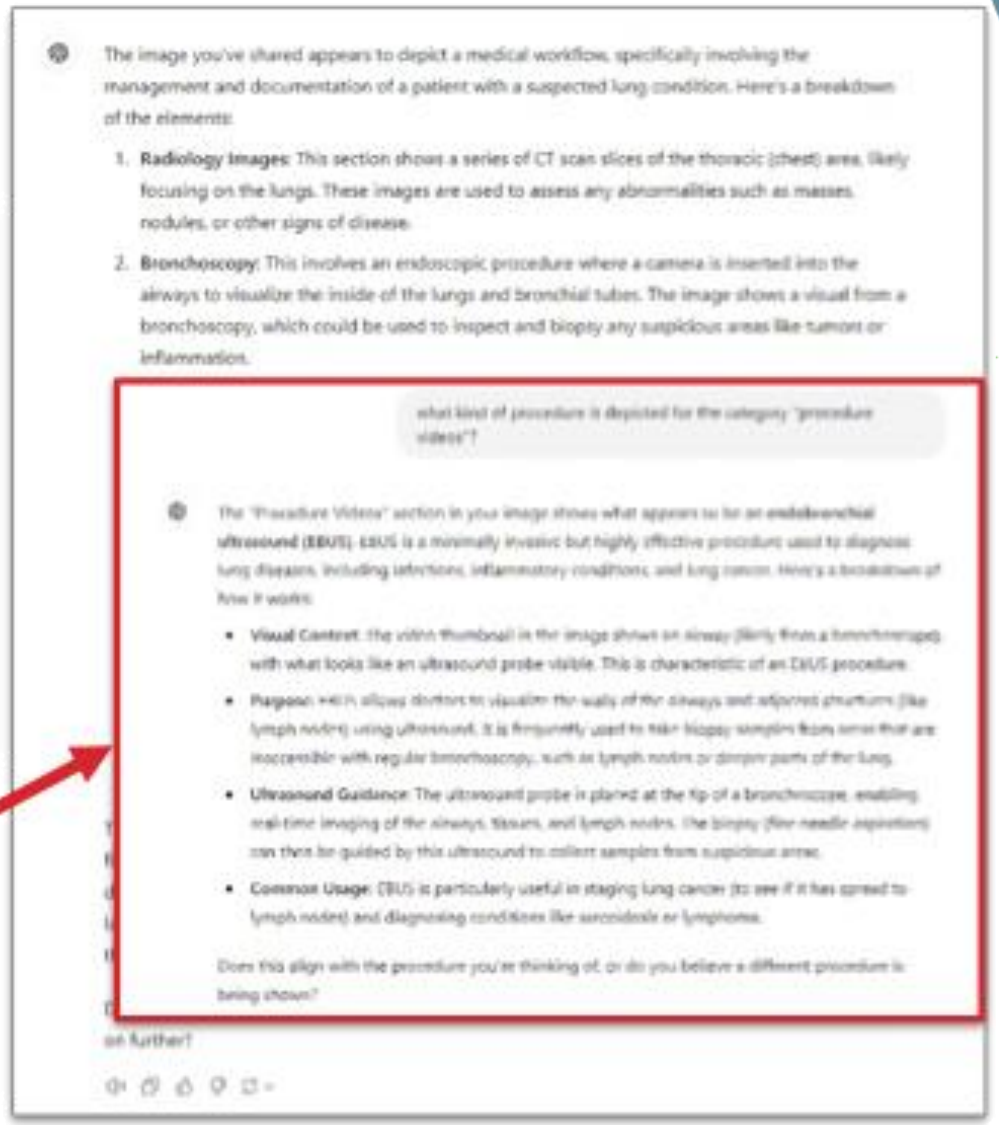


Figure 3. Distributions Representing Smoothed Probability of Changes in Clinical Outcomes (Depression, Anxiety, Weight Concerns, Row-Wise) Postintervention (4 Weeks, Left Column) and at Follow-Up (8 Weeks, Right Column).

The probabilities shown in these plots are derived directly from the CLMMs through predicted probabilities for each possible change score under treatment and control conditions. The treatment group is visualized in blue, and the control group is visualized in red. CLMM denotes cumulative-link mixed model; GAD-Q-IV, Generalized Anxiety Disorder Questionnaire for the *Diagnostic and Statistical Manual of Mental Disorders*, Fourth Edition; PHQ-9, Patient Health Questionnaire 9; and WCS, Weight Concerns Scale.

LLMs have impressive zero-shot abilities





Zero-shot!

LLMs come with limitations

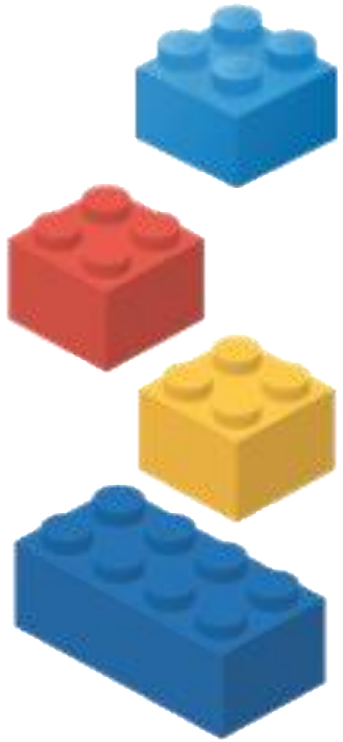
LLMs hallucinate or present factual inaccuracies

LLMs are better reasoning engines, do not use them as knowledge bases

If your model is not hosted locally, you can never be sure about data privacy

Always be sensitive about bias and fairness issues

How to build a great prompt



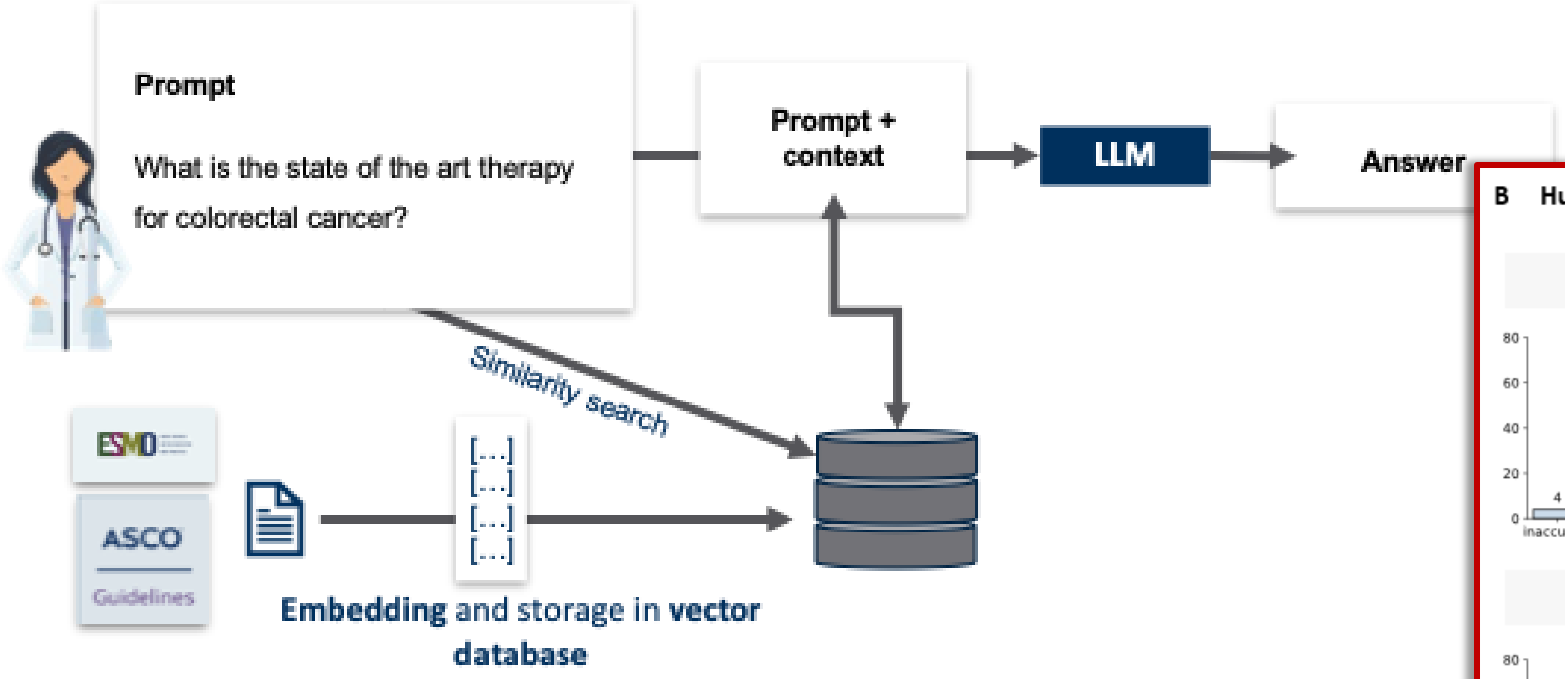
Use examples

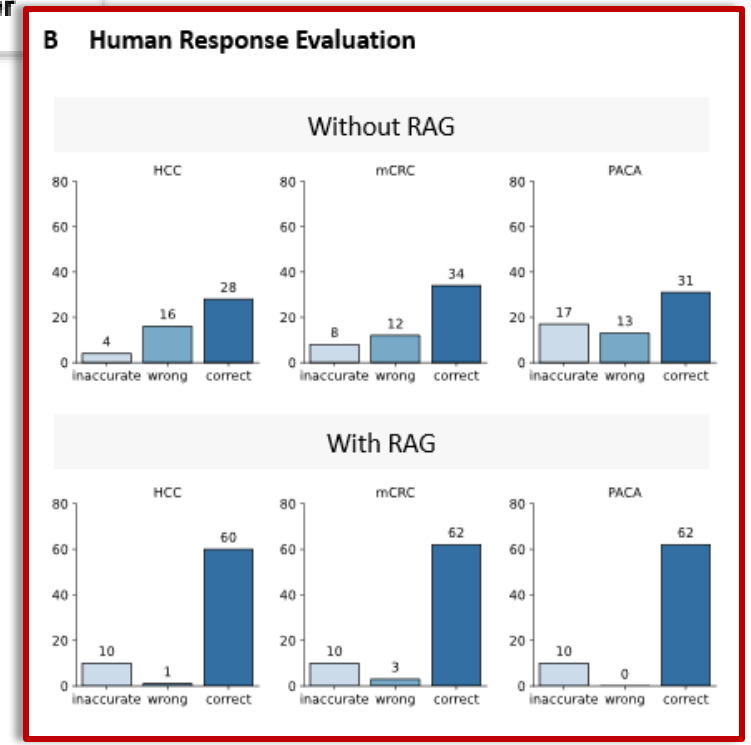
Define style, audience, role

Give structure

Be clear and precise

LLMs can be augmented with information from solid knowledge databases

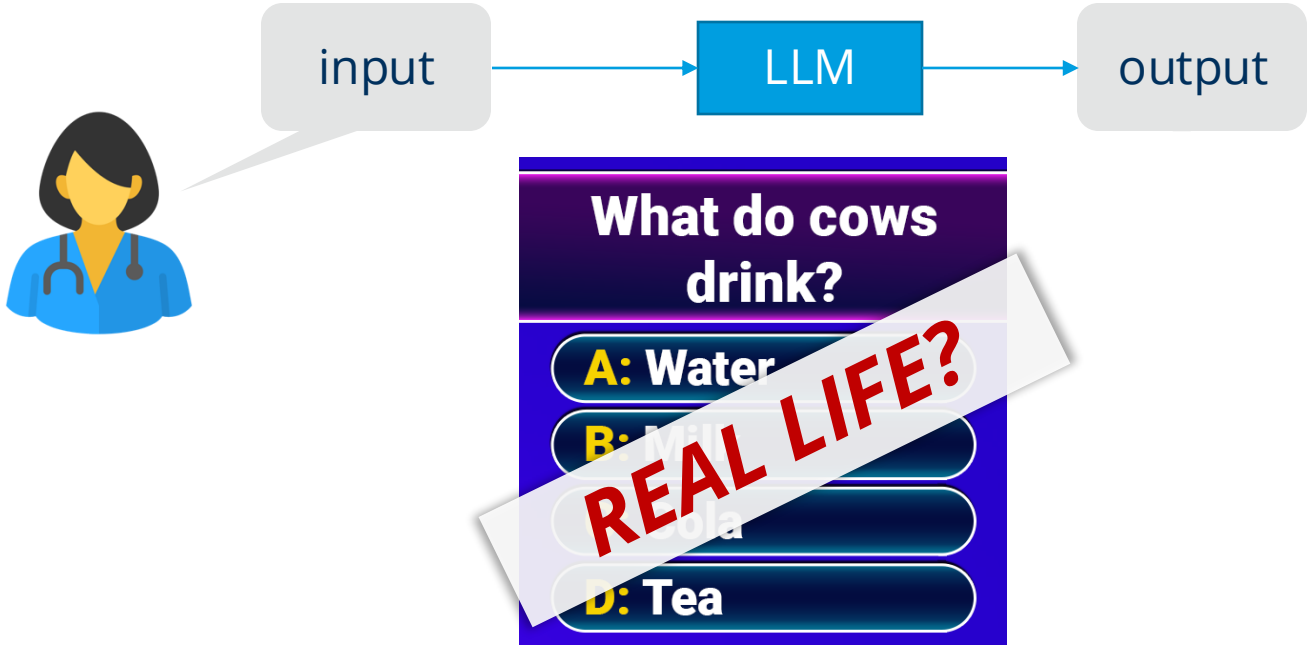




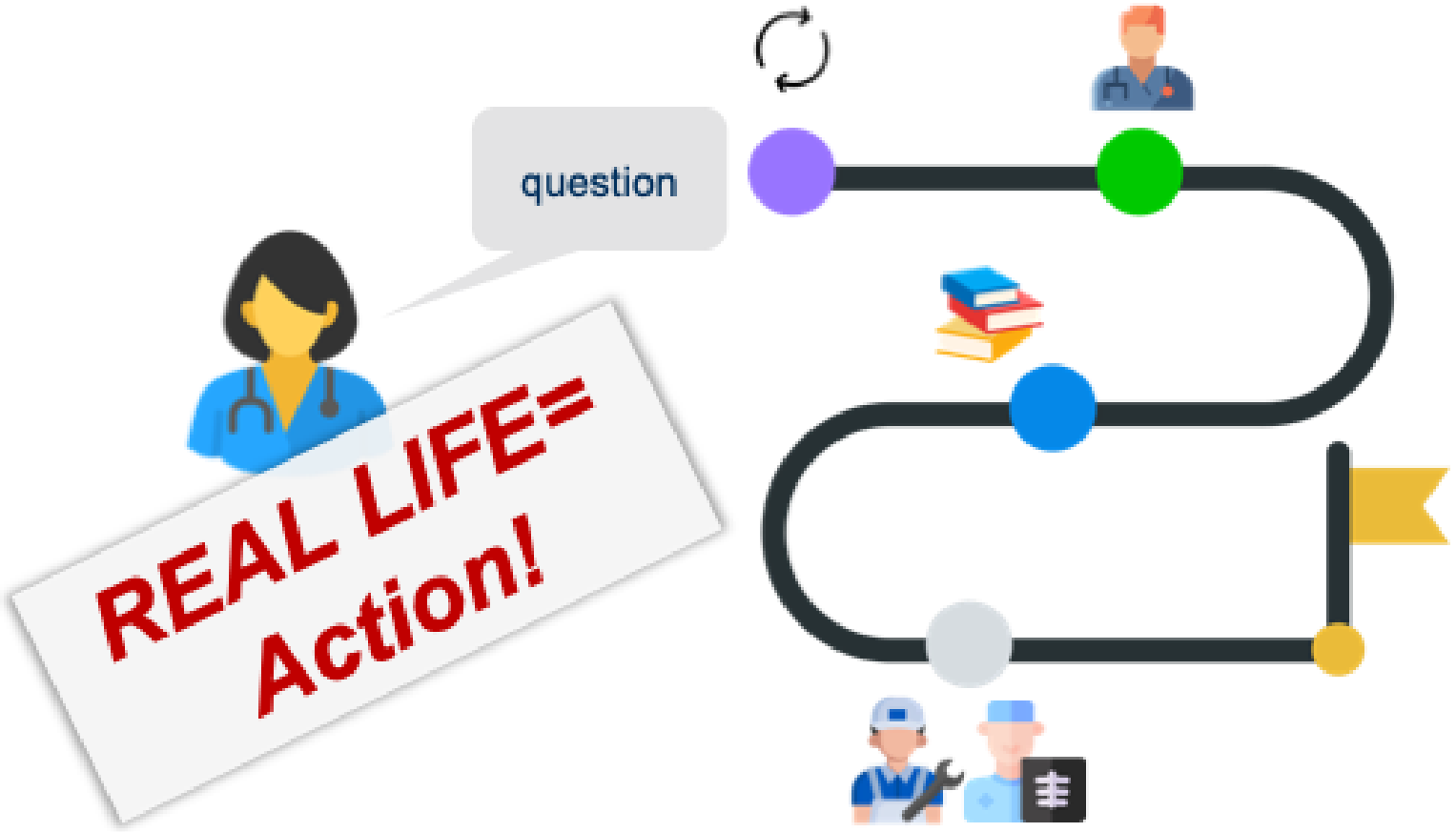
Ferber, D., Wiest, I. C., Wölflein, G., Ebert, M. P., Beutel, G., Eckardt, J. N., ... & Kather, J. N. (2024). GPT-4 for Information Retrieval and Comparison of Medical Oncology Guidelines. NEJM AI, AIcs2300235.

- **Prompt engineering**: clear instructions, structure, examples
- **Use as reasoning engines,** not static knowledge bases
- **Reliable knowledge base**: retrieval, curated corpora, RAG
- **Stepwise reasoning**: "think step by step," chain-of-thought
- **Tool use & multimodality**: calculators, EHRs, APIs, websearch
- **Human-in-the-loop**: expert validation & oversight
- **Domain adaptation**: fine-tuning or instruction tuning
- **Cross-checking**: ensembles, multiple prompts, consistency
- **Safety guardrails**: rules, filters, clinical guidelines
- **Awareness & training**: educate users about risks/limits

The era of agentic AI - from models to AI Agents

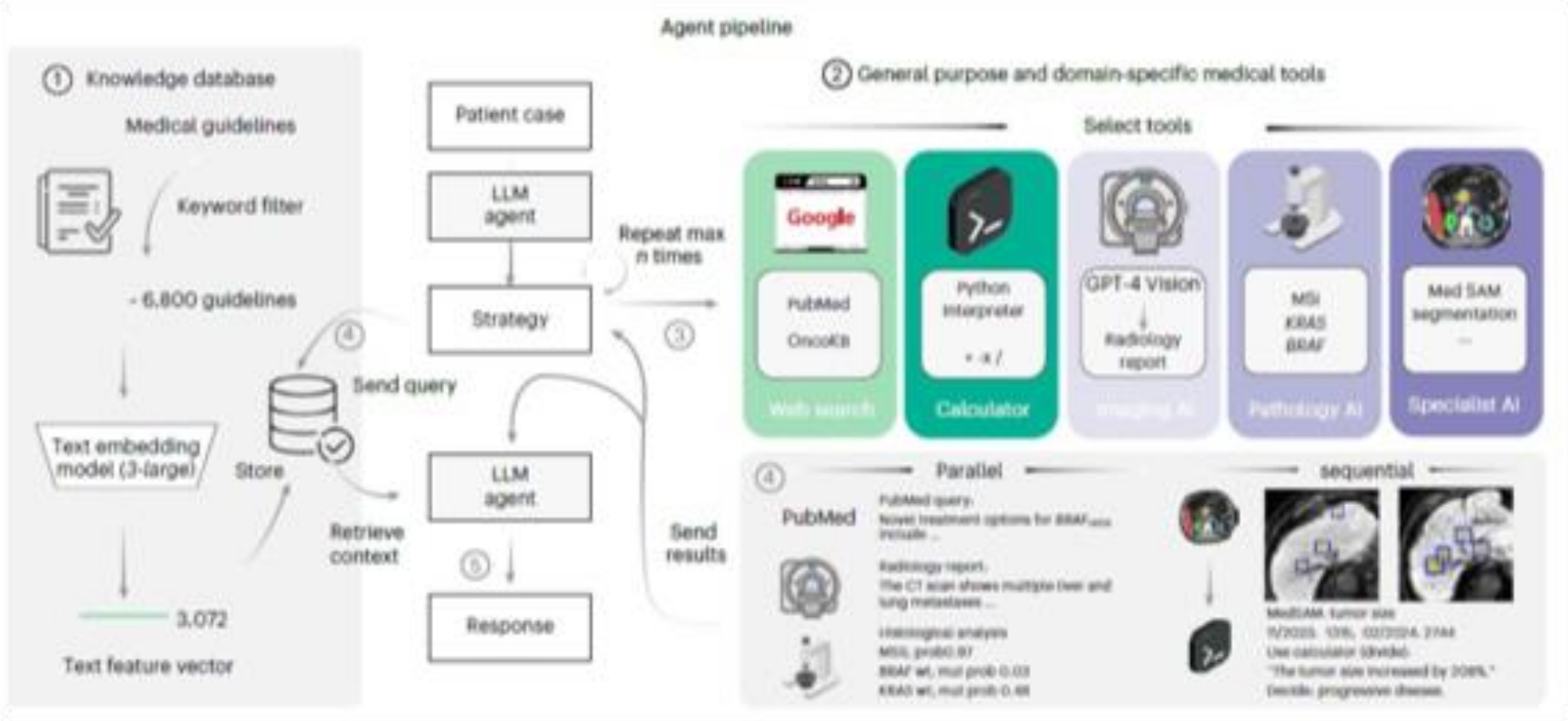


The era of agentic AI - from models to AI Agents



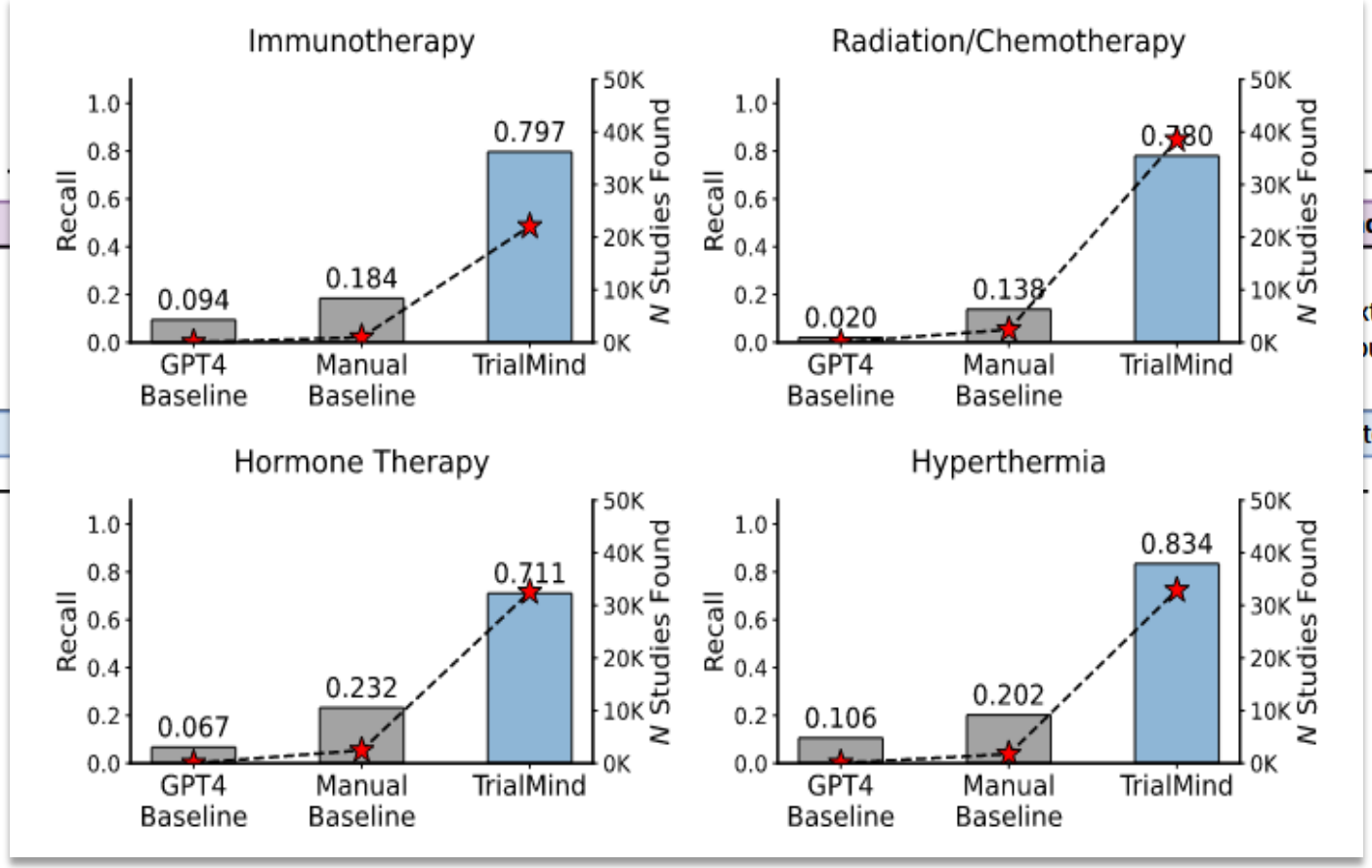
- ✓ Reason
- ✓ Consult experts
- ✓ Knowledge
- ✓ Tools

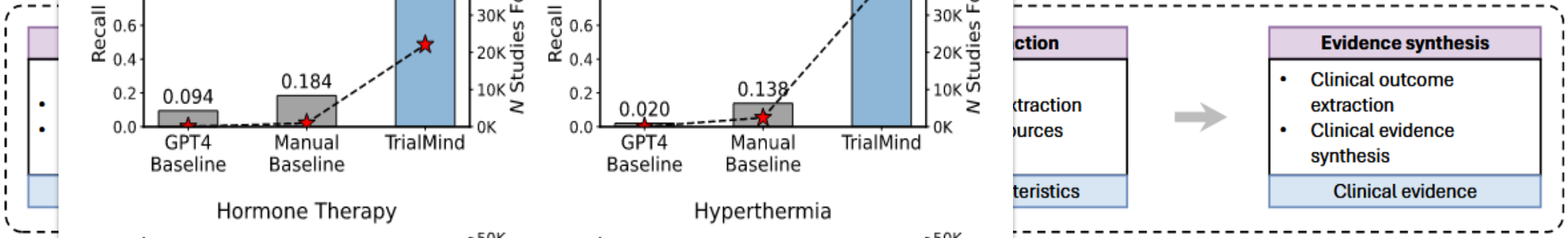
LLM agents for Clinical Decision Support



Ferber, D., El Nahhas, O.S.M., Wölflein, G. et al. Development and validation of an autonomous artificial intelligence agent for clinical decision-making in oncology. Nat Cancer 6, 1337–1349 (2025). https://doi-org.ezproxy.medma.uni-heidelberg.de/10.1038/s43018-025-00991-6

The era of agentic AI - from models to AI Agents

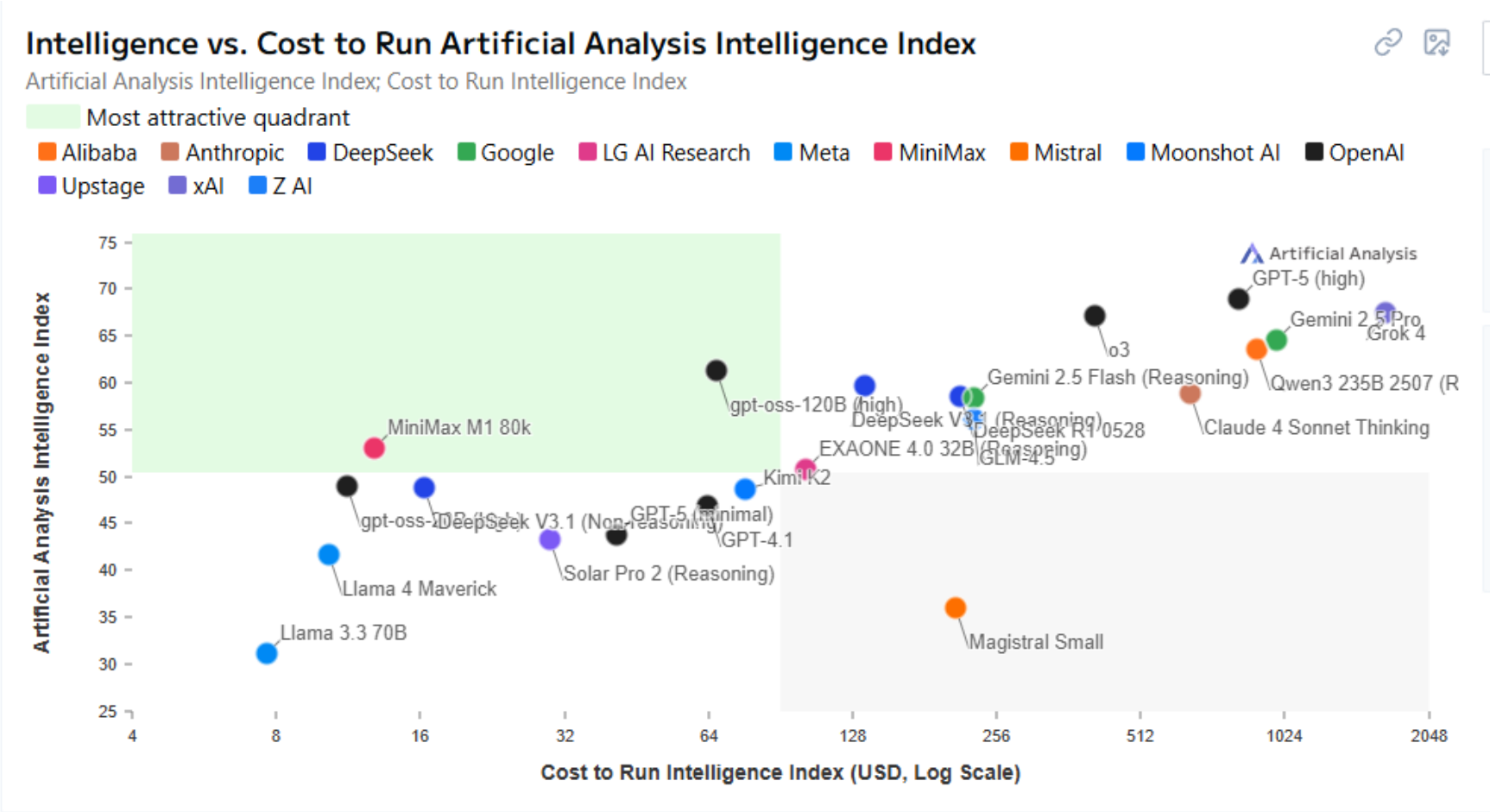




System-designed LLMs outperform humans and simple LLMs

Wang, Zifeng, et al. "Accelerating clinical evidence synthesis with large language models." *npj Digital Medicine* 8.1 (2025): 509.

The era of agentic AI – questions for implementation



https://artificialanalysis.ai/

Model Performance and real world effectiveness might differ

JAMA Network | Open

Original Investigation | Health Informatics

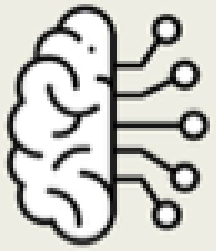
Large Language Model Influence on Diagnostic Reasoning

A Randomized Clinical Trial

Ethan Goh, MBBS, MS[1,2]; Robert Gallo, MD[3]; Jason Hom, MD[4]; et al

INTERVENTION

50 Participants randomized

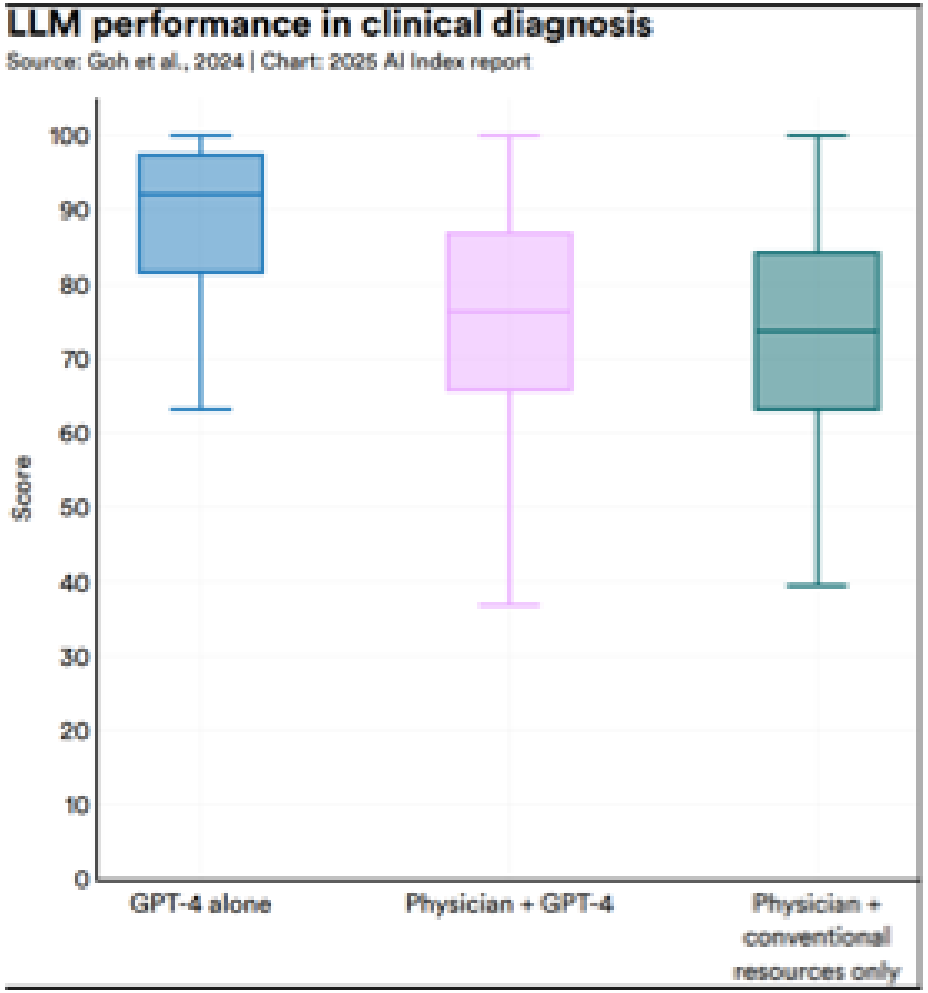


25 Generative artificial intelligence (AI) chatbot

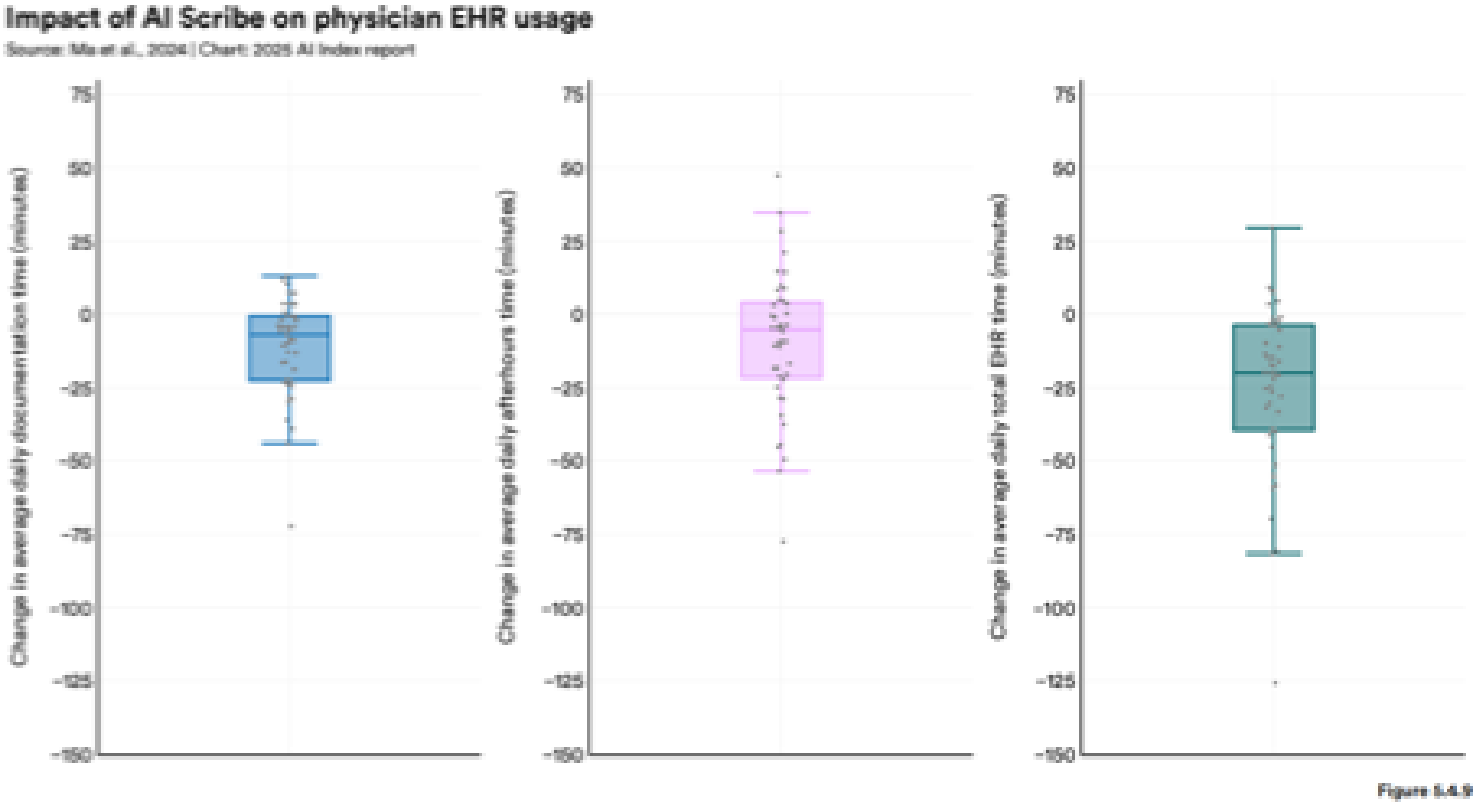
Participants with access to AI chatbot were allocated 60 min to review up to 6 clinical vignettes

25 Conventional resources

Participants with access to conventional online resources such were allocated 60 min to review up to 6 clinical vignettes



Find Use Cases that Matter



AI scribe led to

↓ 30 seconds per note

↓ 20 minutes EHR time per day

↓ ~30% less burden and burnout

Any questions or remarks?

Let’s practice – Python Basics in Jupyter Lab