Large Reasoning Language Models in Clinical Applications

Ben Zhou

Assistant Professor, Arizona State University

https://arc-asu.github.io/

June 2026

Language Models

- Language Models (LMs) are models that attempt to predict word distributions given a context.

Context:

> *"I like fruits, so I ate two"*

Next Word Candidate	Probability (%)
apples	32.5%
bananas	24.0%
mangoes	10.3%
oranges	8.7%
pears	6.5%
strawberries	4.2%
grapes	3.1%
slices	2.0%
sandwiches	1.4%
cookies	1.0%

A good next-word prediction model encompasses logic and reason.

Model: OpenAI GPT-4o

Large Language Models

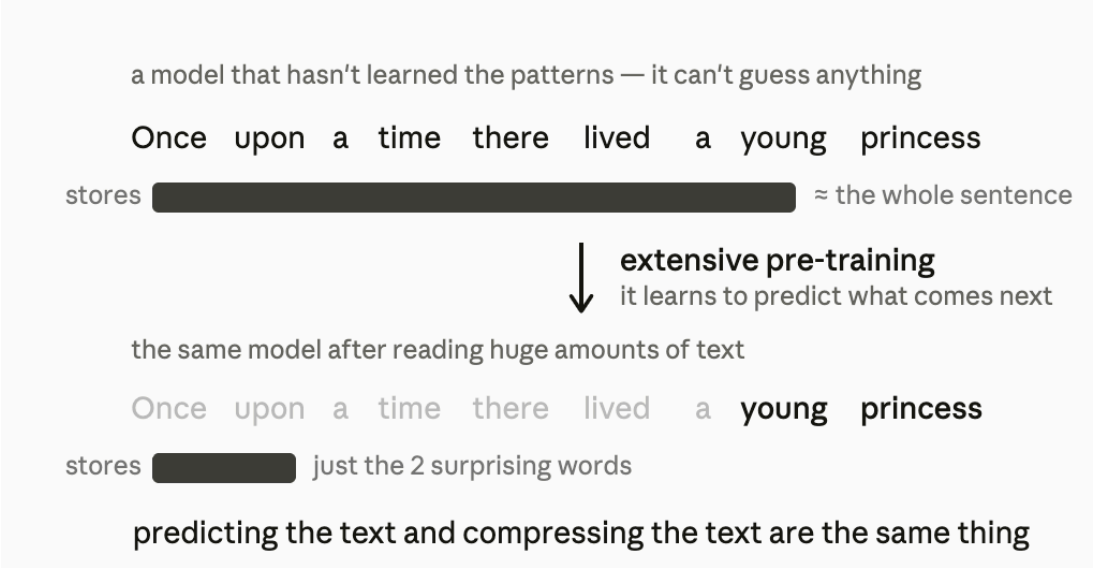
- Language models have existed for many decades.
  - First sequential probabilistic modeling: ~1906
  - First “language model”: Claude Shannon, 1948
- Modern Large Language Models (LLMs) differ in a few ways:
- 1. Larger models based on transformers (2017)
  - LLMs usually have 0.1B to ~1000+B number of parameters
    - Progressed from 0.1B to 1000B in just 8 years.
  - Very large-scale computation
    - Partial reason of the memory price hikes and stock prices on Nvidia/Micron

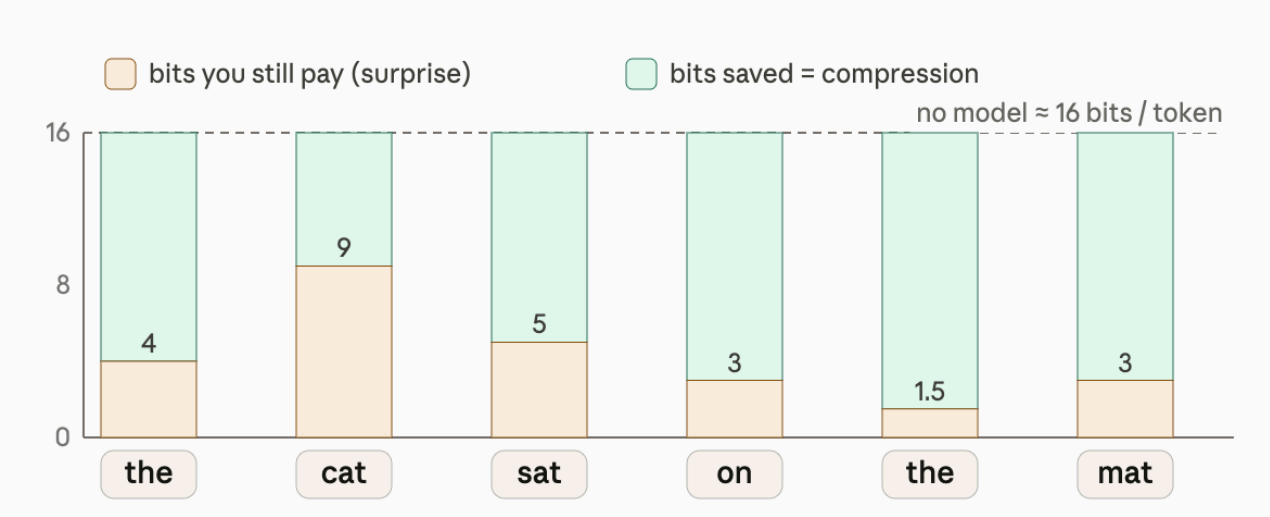
Large Language Models

- Language models have existed for many decades.
  - First sequential probabilistic modeling: ~1906
  - First “language model”: Claude Shannon, 1948
- Modern Large Language Models (LLMs) differ in a few ways:
- 2. LLMs are pre-trained on a huge amount of raw texts
  - From <1B tokens (2018) to 36T tokens (Qwen3, 2025)
  - Commercial models may use much more data

What does it mean to have bigger models?

- A successful pre-training equals “compressing” the raw texts used in training into the model parameters.
  - Pre-training: use raw texts and training models to predict the next word





What does it mean to have bigger models?

- A successful pre-training equals “compressing” the raw texts used in training into the model parameters.
- Implication: when the model is “compressing” raw texts internally, it is also compressing the “knowledge” and “patterns” encoded in raw texts – which documents many things.
  - This is the key assumption from LLMs to “AI”.

Mid-training

- Even though pre-training encodes knowledge into the model, we still need to “translate” that knowledge into a form that we can use the most effectively.
- Mid-training uses labeled “input->output” pairs to tell the model what to say under specific instructions.
  - The training method is also known as Supervised Fine-tuning (SFT).

Prompt:

Copy

```
Instruction: Convert the following sentence into passive voice.
Input: The chef prepared a delicious meal.
```

Response:

Copy

```
A delicious meal was prepared by the chef.
```

Multi-modal Mid-training

- At mid-training stages, we can add on other “modalities” to the model, such as images, videos and sound.

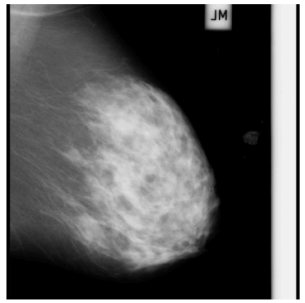
Problem 2: What content appears in this image?

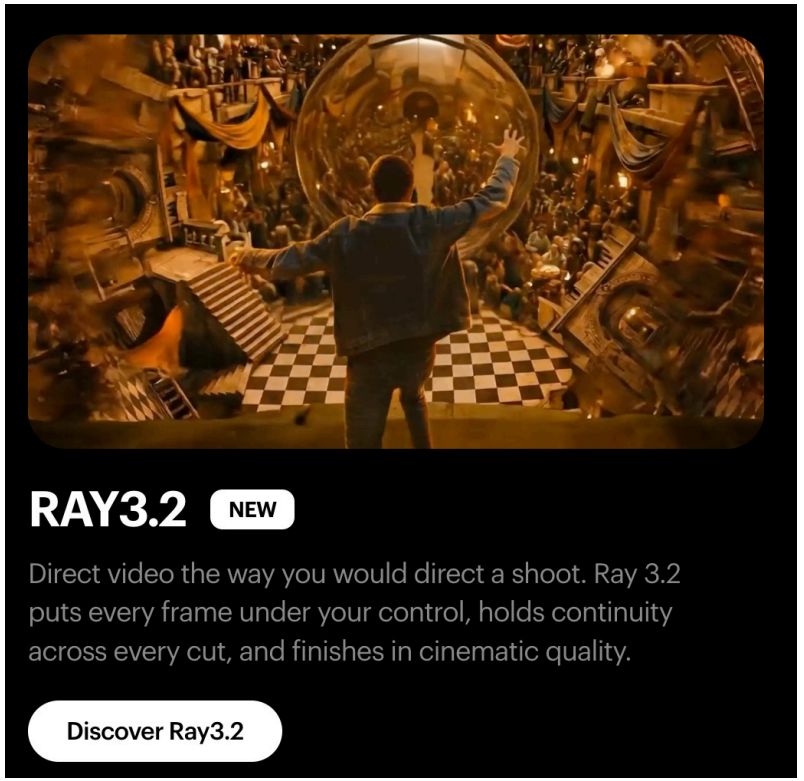
A) Cardiac tissue, B) Breast tissue, C) Liver tissue, D) Skin tissue

MedVLM-R1 Output:

<think> The image appears to be a mammogram, which is a type of X-ray used to detect abnormalities in the breast tissue. The image shows the breast tissue with various densities and patterns, which are typical of mammograms. </think>

<answer>B</answer> **Groundtruth Answer: B**

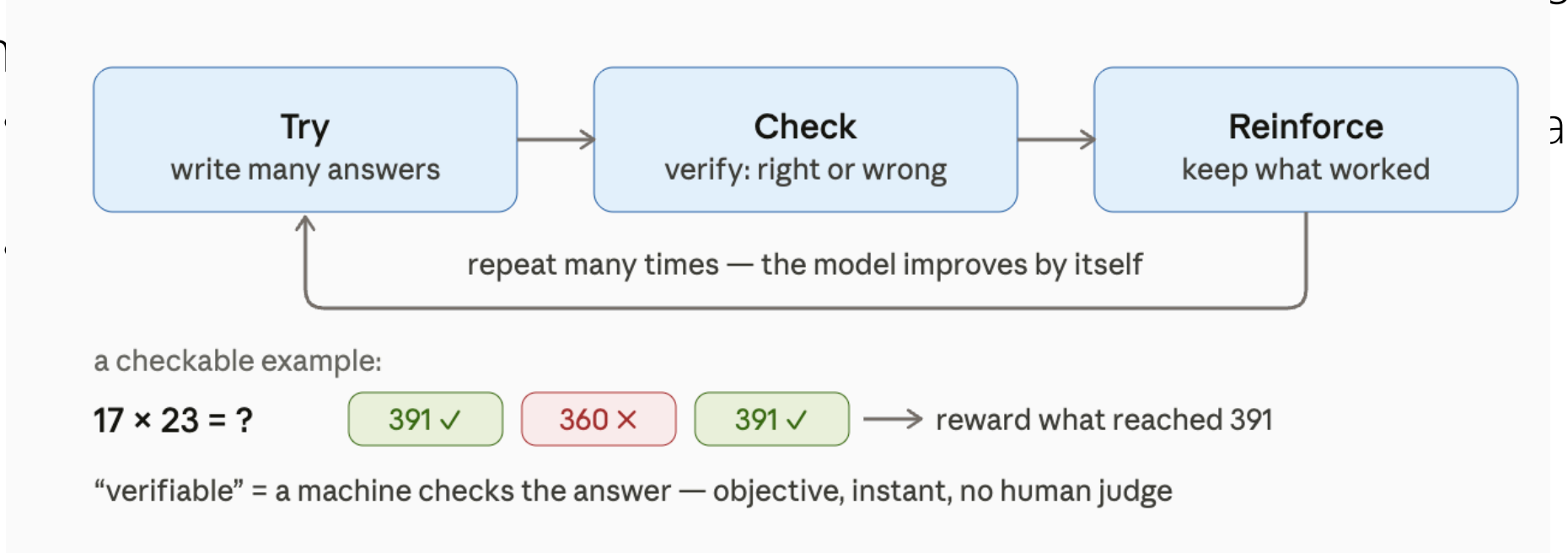




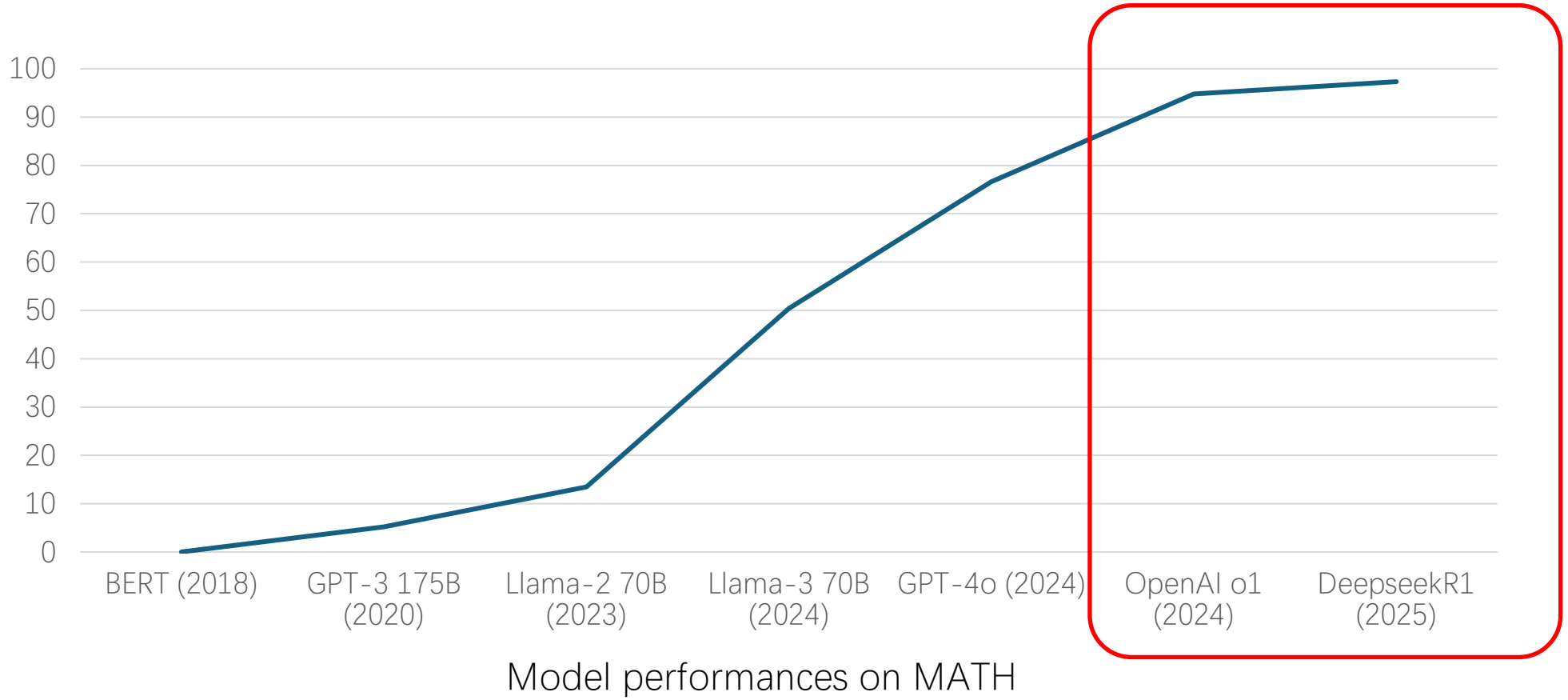
Now multi-modal LLMs can understand and generate images, videos and sounds, by clever mid-training that maps the knowledge and skills compressed in raw texts to representations of other modalities.

Post-training

- After pre-training and mid-training, models are very good at chatting with us, following our instructions, and answer questions.
- However, a big problem in mid-training is that we always tell the m



Post-training enabled “Reasoning” LLMs



Model performances on MATH

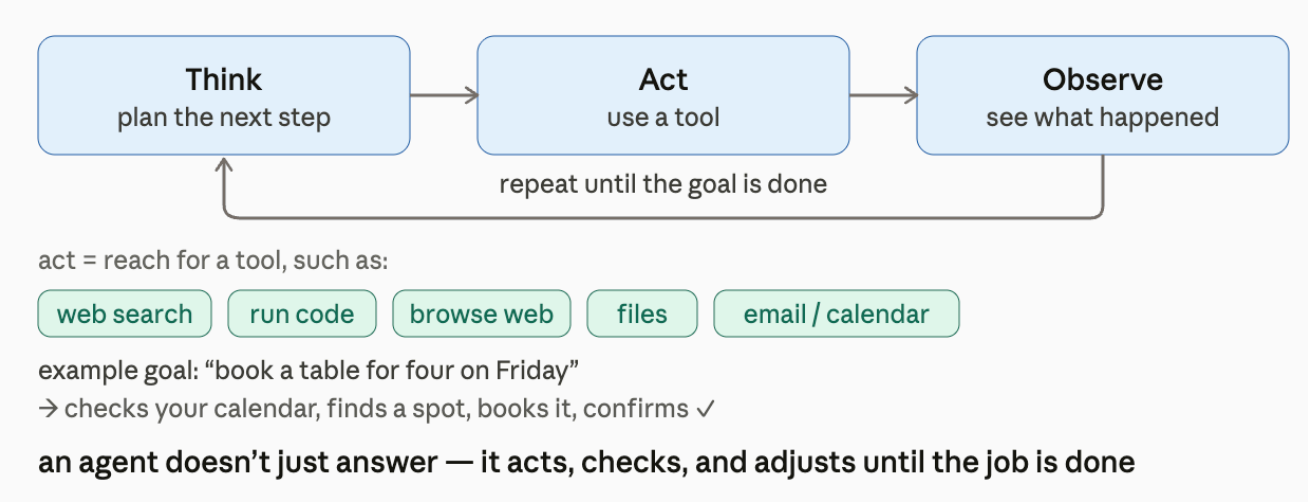
Reasoning LLMs and the Future

- For a very long time, LLMs were only able to solve tasks that humans can solve too, just faster and cheaper.
- Until recently

A problem that mathematicians have actively worked on for decades. Solved by an LLM.

RL has enabled agents, too

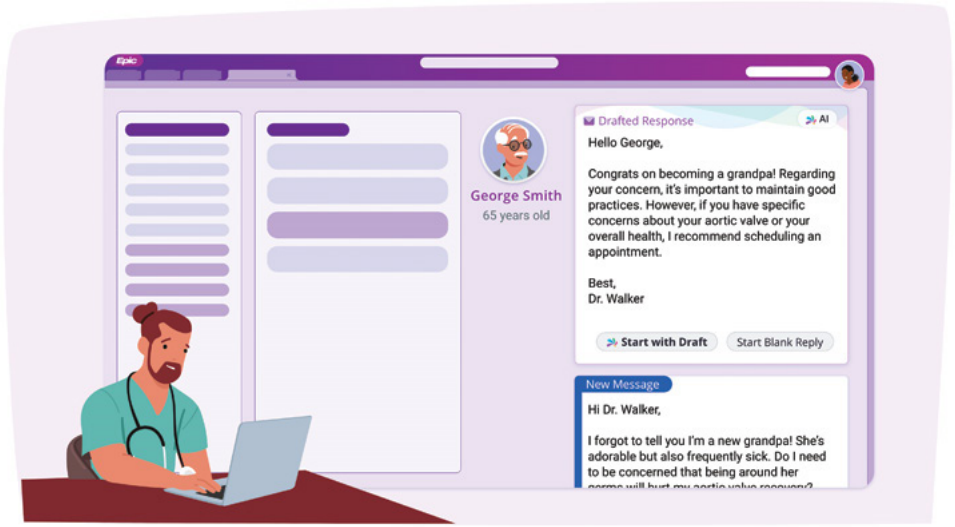
- What is an agent?
  - A LLM that iteratively calls “tools (MCPs, skills)”, reflect/reason on tool-call results, and plan for the next action.
- In clinical settings, tools can be web search, EHR search, escalations, filling forms, chat with patients, etc.
  - Anything that a LLM cannot do by itself can be achieved via tools.



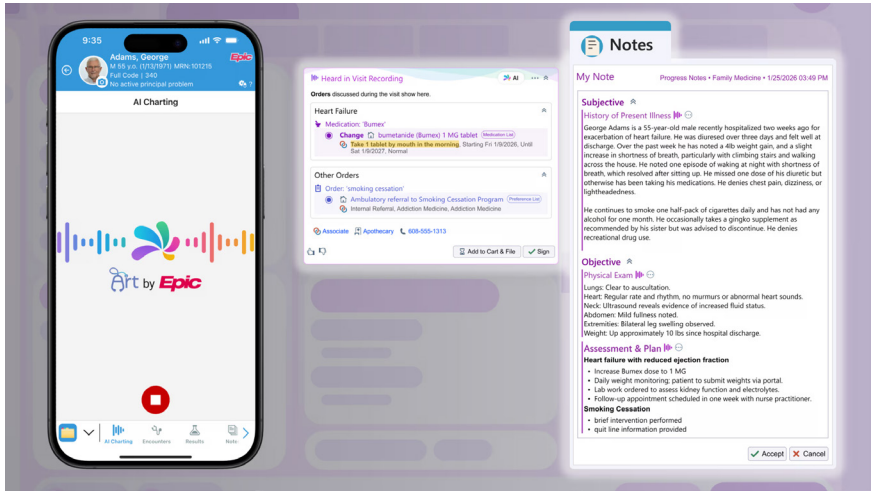
RL is key in agents: it is hard to annotate tool call processes for all tasks; RL allows models to extrapolate to many tasks.

Where is AI in Clinical Settings?

- Agents, multimodal applications and deep reasoning enabled by LLMs are becoming everywhere in clinical settings.
- Simple ones such as



Message drafting



Auto charting/scribing

Image sources: Epic

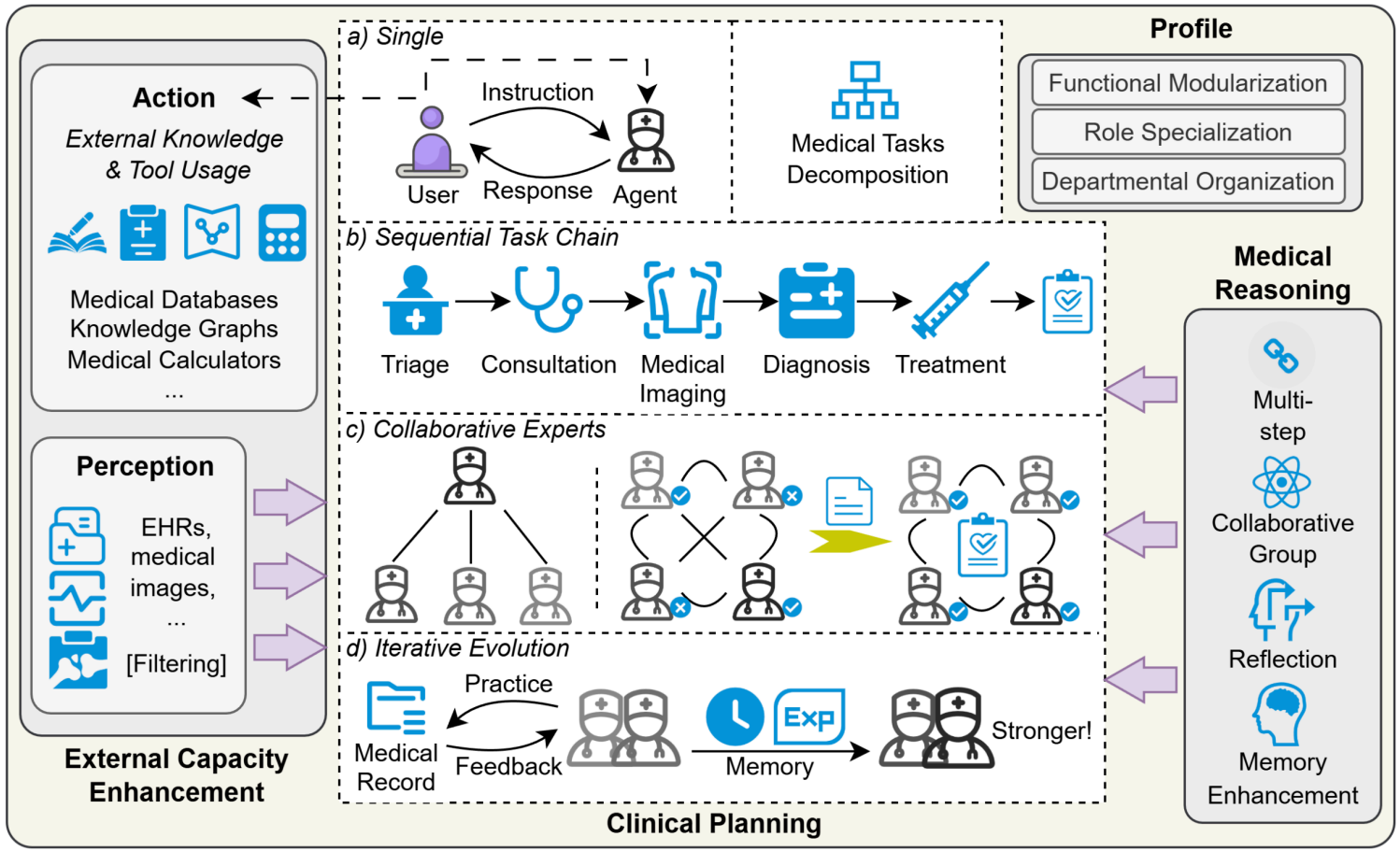
Radiology

- Liang et al. 2026: AI outperforms radiologist on **Breast MRI BI-RADS 4 lesion classification**.
- Winkel et al. 2025: As **Mammography as second reader,** AI had higher cancer detection than single human reading.
- Anderson et al. 2024 (relatively old tech): AI performed roughly on par with radiologists on **Chest X-ray abnormality detection.**

Disclaimer: I did not read these papers, and I am blindly trusting the reviews of the medical journals that these papers are published in, which, in many cases, may be wrong.

Clinical Agents

- For triaging, diagnosis, treatment, monitoring, discharging.



Source: Wang et al. A Survey of LLM-based Agents in Medicine: How far are we from Baymax?

Agentic Self Evolution

- A particular hype on agents in complex clinical tasks is that many people believe that the models can “self-evolve”.
  - Even though current agents are not perfect, we can still deploy them in the clinical setting, quietly observing and improving via RL.

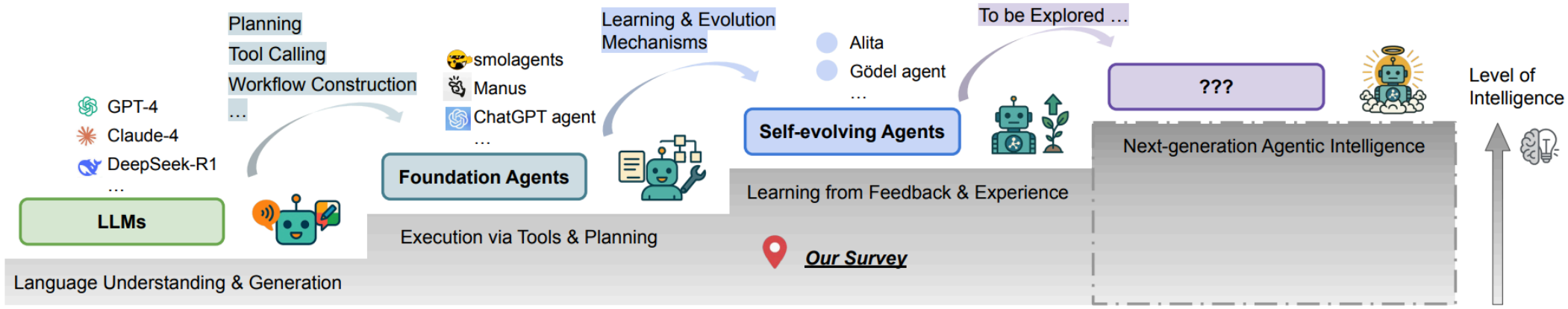


Figure from “A Survey of Self-Evolving Agents”

Where are we in clinical AI journey?

- If LLMs and AI agents are so good, why are we not seeing concrete studies that show significant improvements in
  - Clinician well-being
  - Patient care delivery time
  - Care quality
  - Or major workforce shifting in the clinical field?


AI

Amazon Web Services CEO says half of white-collar jobs may 'change' due to AI — but it won't be a 'wipe-out'

By Natalie Musumeci + Follow

Oracle lays off 21,000 employees in just 12 months due to AI adoption and costly AI infrastructure ambitions — says layoffs will continue as internal AI deployment grows

News By Etiido Uko published yesterday

This is not the case in other fields.

Reason 1: Accountability

- Unlike many other jobs that AI is shifting, clinical workflows often require rigorous verifiability, interpretability, and auditability.
  - This is not necessarily all about accuracy: many AI agents are already better than average human clinicians in narrow tasks.
  - This is more about “accountability”: who takes the blame when a mistake is made?
- We cannot blame the model, so the only solution is to provide guarantees on model performance.
  - But this is very challenging, or nearly impossible in certain scenarios.

Reason 2: Controllability

- Once we have an AI deployed in clinical settings with some level of automation, we must be able to control model's behavior.
- However, AI can go rogue, not because they are "bad", but because objectives often create conflicts.
- At the same time, agents may be "attacked" by malicious parties.

Anthropic: models may send blackmails if they feel like that's the right way to solve a task.

Reason 3: Policy

- There are many ethical and political concerns that lead to policies preventing AI deployment in clinical settings.
- For example, if on certain tasks, we can show AI is on par with human. AI presumes no liability. What will a patient choose?
- With older models (GPT-4 equivalent):
  - Takita et al. 2025: on 83 diagnostic cases, AI decisions has no significant difference with physicians overall
  - Goh et al. 2025: on certain patient-care cases, LLM scored significantly higher than physicians, and on-par with physicians + LLM augmentation.

Reason 3: Policy

- There are many ethical and political concerns that lead to policies preventing AI deployment in clinical settings.
- For example, if on certain tasks, we can show AI is on par with human. AI presumes no liability. What will a patient choose?
- With newer models (GPT-5 equivalent):
  - Brodeur et al. 2026: AI matched or exceeded physicians across several diagnosis and management-reasoning tasks, including real emergency-department data.
  - Google AMIE: AI has higher diagnostic accuracy and better performance on many consultation-quality axes, stronger plan precision/guideline alignment

Reason 3: Policy

- There are many ethical and political concerns that lead to policies preventing AI deployment in clinical settings.
- For example, if on certain tasks, we can show AI is on par with human. AI presumes no liability. What will a patient choose?
  - The real difference: For patients, full AI^ is much cheaper and faster.
- Timewise: in the U.S., patients on average wait for 23.5 days for family medicine, and 40+ days for some specialties*
  - AI has no wait

*2025 Survey of Physician Appointment Wait Times and Medicare and Medicaid Acceptance Rates
^Here full AI still requires human operators on machines/tests, or robotic developments that we do not currently have.

Reason 3: Policy

- There are many ethical and political concerns that lead to policies preventing AI deployment in clinical settings.
- For example, if on certain tasks, we can show AI is on par with human. AI presumes no liability. What will a patient choose?
  - The real difference: For patients, AI is much cheaper and faster.
- Cost-wise: Dermatologists see ~2,300 cases and are compensated $350K per year. Per case: ~$150.^, cardiologists $250
  - SOTA AI (e.g., GPT-5.5) on a complex case costs $20*

^Very rough, non-scientific estimation, from estimated total cases, divided by total licensed specialty clinicians, calculated with Bureau of labor statistics average salary.

*Assuming GPT 5.5, 2M input tokens and 200k output tokens. Estimated from running actual end-to-end cases. This is the price paid to OpenAI, not actual OpenAI cost.

Reason 3: Policy

- There are many ethical and political concerns that lead to policies preventing AI deployment in clinical settings.
- For example, if on certain tasks, we can show AI is on par with human. AI presumes no liability. What will a patient choose?
  - For patients, AI is much cheaper and faster.
  - **Many** people may choose AI if this is ever allowed.
- On many tasks, the tech is already real.
- Ethical-wise, not really. However, people tolerate ethical imperfections when the alternative is no care.

Are AI-assistants Actually Assisting?

- Given that we cannot yet deploy AI models with semi or full autonomy, most AI applications today in clinical settings are to assist human clinicians.
  - Some applications are not saving time but rather increase time and error.

DIVE BRIEF

AI scribe adoption linked to modest reductions in EHR, documentation time: study

Clinicians' use of an AI scribe was associated with 13 fewer minutes each day in the EHR and 16 fewer minutes on documenting patient care, according to the research published in JAMA.

Published April 2, 2026

Can incorrect artificial intelligence (AI) results impact radiologists, and if so, what can we do about it? A multi-reader pilot study of lung cancer detection with chest radiography

Michael H Bernstein [1,2,3,✉,#], Michael K Atalay [1,3,#], Elizabeth H Dibble [1], Aaron W P Maxwell [1,3], Adib R Karam [1], Saurabh Agarwal [1], Robert C Ward [1], Terrance T Healey [1], Grayson L Baird [1,2,3]

Author information · Article notes · Copyright and License information

PMCID: PMC10235827 PMID: 37266657

Costs: Is AI Actually Cheaper than Human?

- AI Infrastructure is getting much more expensive. Sometimes it is cheaper to hire humans to do certain tasks.
- Probably not yet the case in clinical fields, but AI will force industries to re-evaluate when to hire humans v. when to use AI.

Newsletter Artificial Intelligence

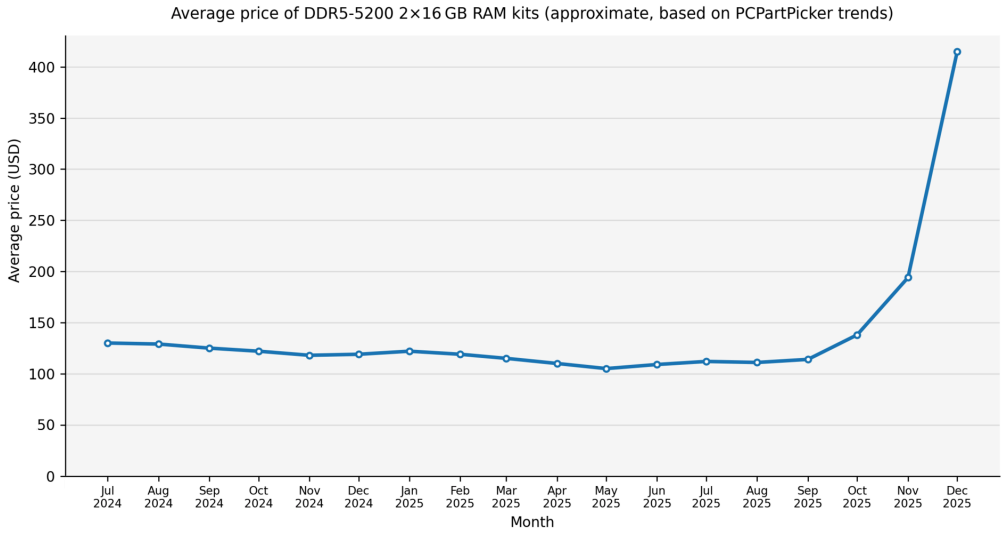
Artificial Intelligencer: Wall Street can't decide what to think about AI

Reuters Artificial Intelligencer

By **Deepa Seetharaman**

Sign up for newsletter

June 24, 2026 1:20 PM MST · Updated 3 hours ago



Source: Wikipedia

Thank you!

- LLMs, multi-modal LLMs, and agents
- Current applications of AI in clinical fields.
- Discussed a few points on the future of clinical AI.
- All the points discussed today are somewhat related to my research interests, feel free to connect and chat more!