

Agents, Reasoning, and LLMs In Healthcare

Mar. 17, 2025

Presenter: Guangzhi Xiong



UNIVERSITY *of* VIRGINIA

Presentation Contents

- Surveys of Medical LLMs
 - A Survey on Medical Large Language Models: Technology, Application, Trustworthiness, and Future Directions
 - A Survey of Large Language Models for Healthcare: from Data, Technology, and Applications to Accountability and Ethics
- Agentic AI in Healthcare
 - The rise of agentic AI teammates in medicine
- Reasoning Models for Medicine
 - From Medprompt to o1: Exploration of Run-Time Strategies for Medical Challenge Problems and Beyond

References

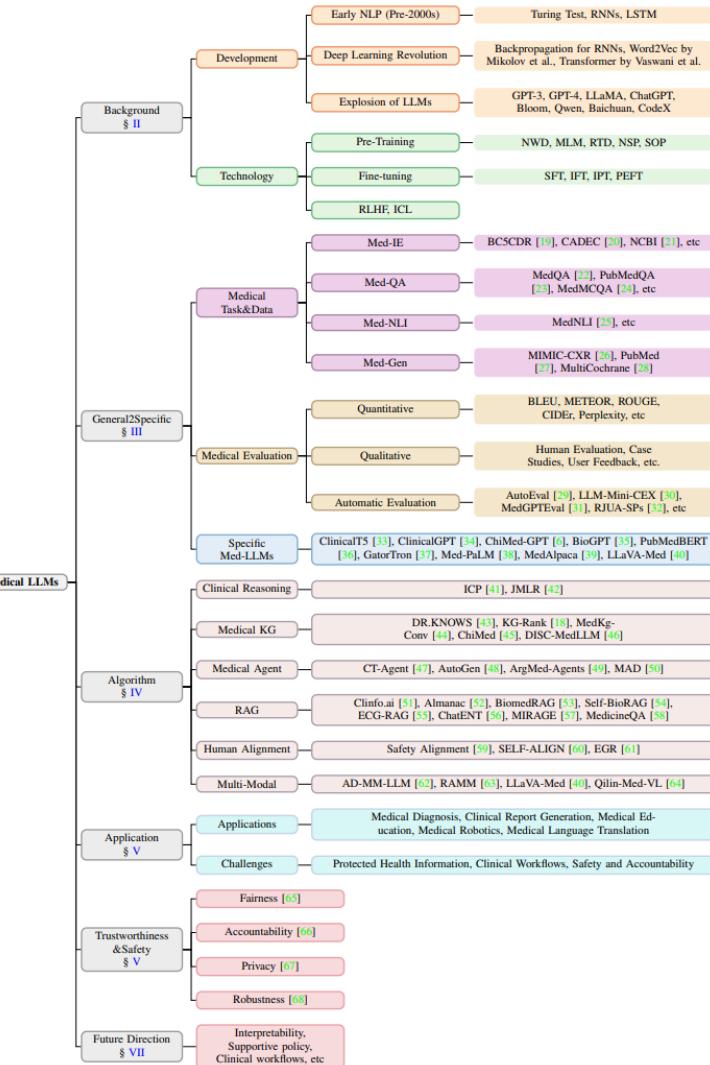
- Zou, J., & Topol, E. J. (2025). The rise of agentic AI teammates in medicine. *The Lancet*, 405(10477), 457.
- Frey, N. C., Hotzel, I., Stanton, S. D., Kelly, R. L., Alberstein, R. G., Makowski, E. K., ... & Gligorijevic, V. (2025). Lab-in-the-loop therapeutic antibody design with deep learning. *bioRxiv*, 2025-02.
- Tang, X., Zou, A., Zhang, Z., Li, Z., Zhao, Y., Zhang, X., ... & Gerstein, M. (2024, August). MedAgents: Large Language Models as Collaborators for Zero-shot Medical Reasoning. In *Findings of the Association for Computational Linguistics ACL 2024* (pp. 599-621).
- Kim, Y., Park, C., Jeong, H., Chan, Y. S., Xu, X., McDuff, D., ... & Park, H. (2024). Mdagents: An adaptive collaboration of llms for medical decision-making. *Advances in Neural Information Processing Systems*, 37, 79410-79452.
- Nori, H., Usuyama, N., King, N., McKinney, S. M., Fernandes, X., Zhang, S., & Horvitz, E. (2024). From medprompt to o1: Exploration of run-time strategies for medical challenge problems and beyond. *arXiv preprint arXiv:2411.03590*.
- Zhong, T., Liu, Z., Pan, Y., Zhang, Y., Zhou, Y., Liang, S., ... & Liu, T. (2024). Evaluation of openai o1: Opportunities and challenges of agi. *arXiv preprint arXiv:2409.18486*.
- Nori, H., Lee, Y. T., Zhang, S., Carignan, D., Edgar, R., Fusi, N., ... & Horvitz, E. (2023). Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*.
- Liu, L., Yang, X., Lei, J., Liu, X., Shen, Y., Zhang, Z., ... & Ren, K. (2024). A survey on medical large language models: Technology, application, trustworthiness, and future directions. *arXiv preprint arXiv:2406.03712*.
- He, K., Mao, R., Lin, Q., Ruan, Y., Lan, X., Feng, M., & Cambria, E. (2025). A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *Information Fusion*, 102963.

A Survey on Medical Large Language Models: Technology, Application, Trustworthiness, and Future Directions

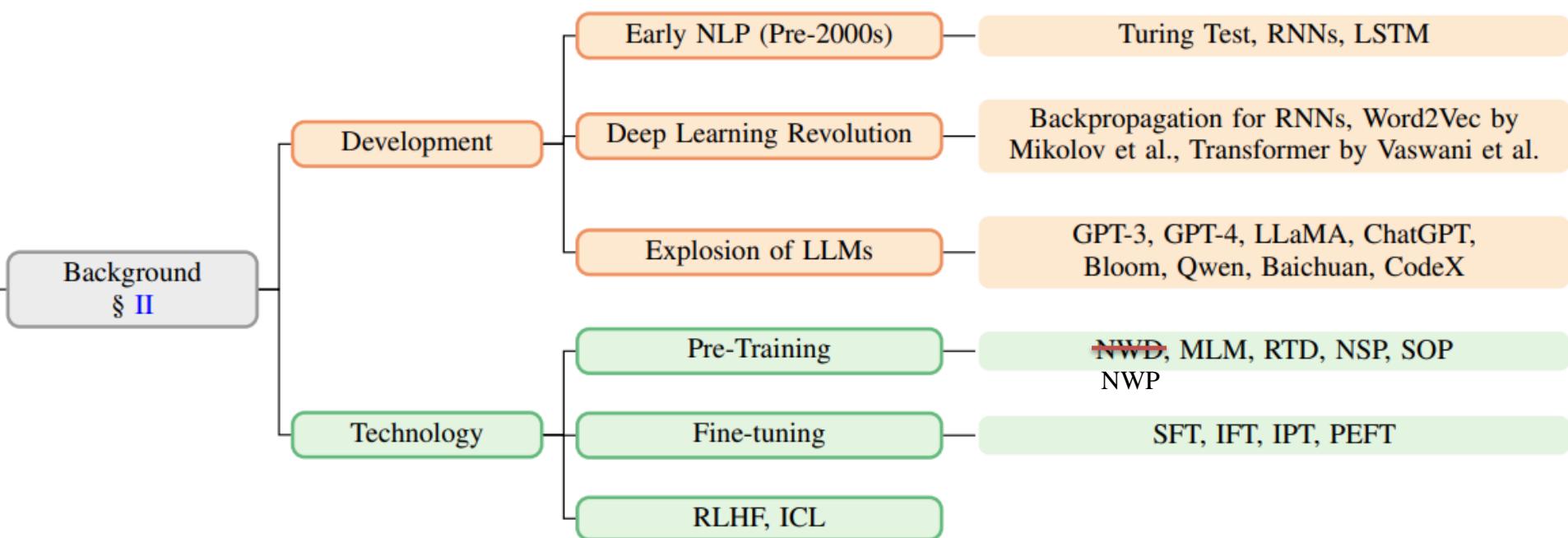
LIU ET AL.
ANT GROUP, ZJU & CUHK
DEC 2024

Organization of the Survey

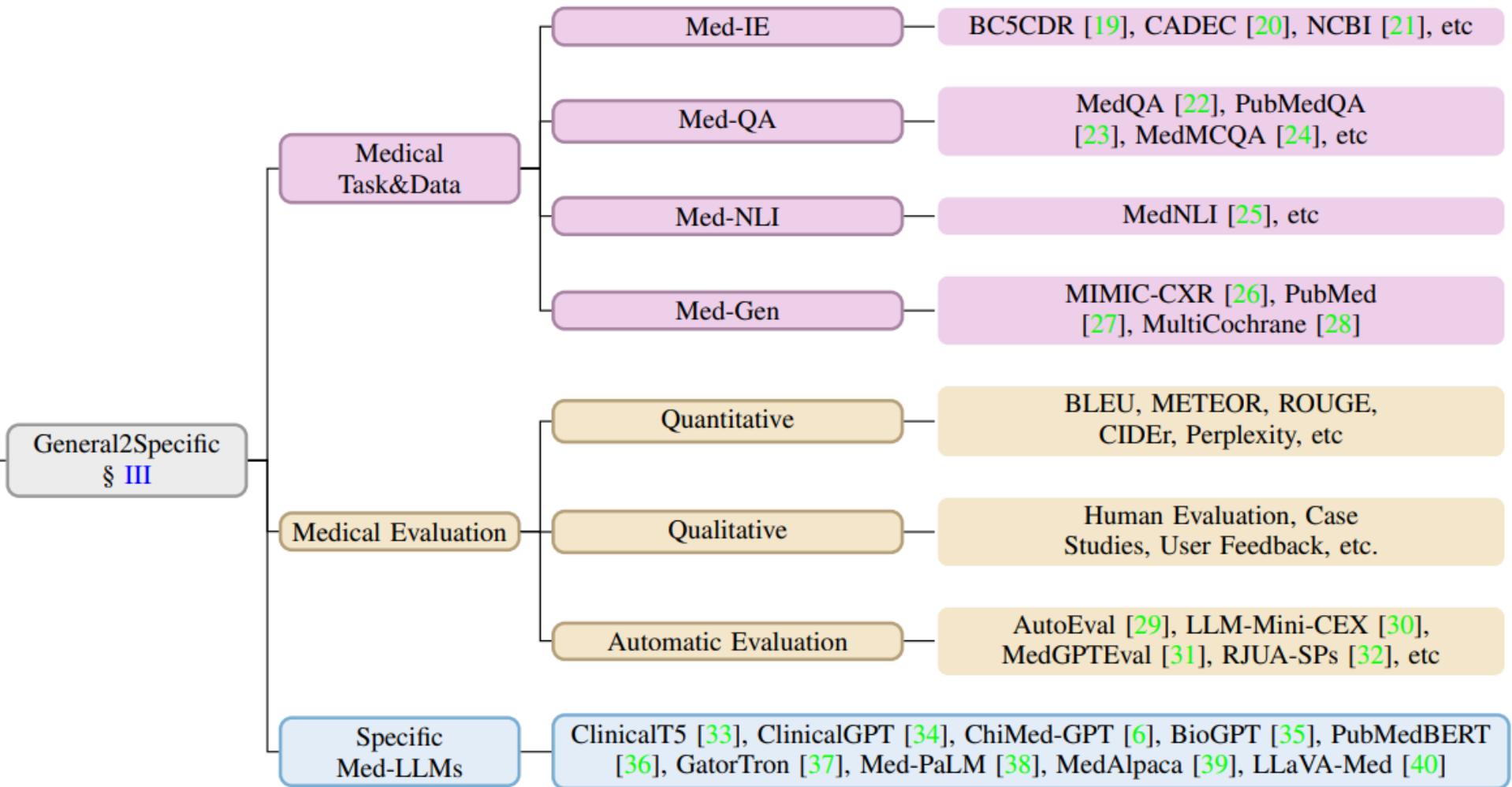
- Background
- General2Specific
- Algorithm
- Application
- Trustworthiness & Safety
- Future Direction



Background & Technology

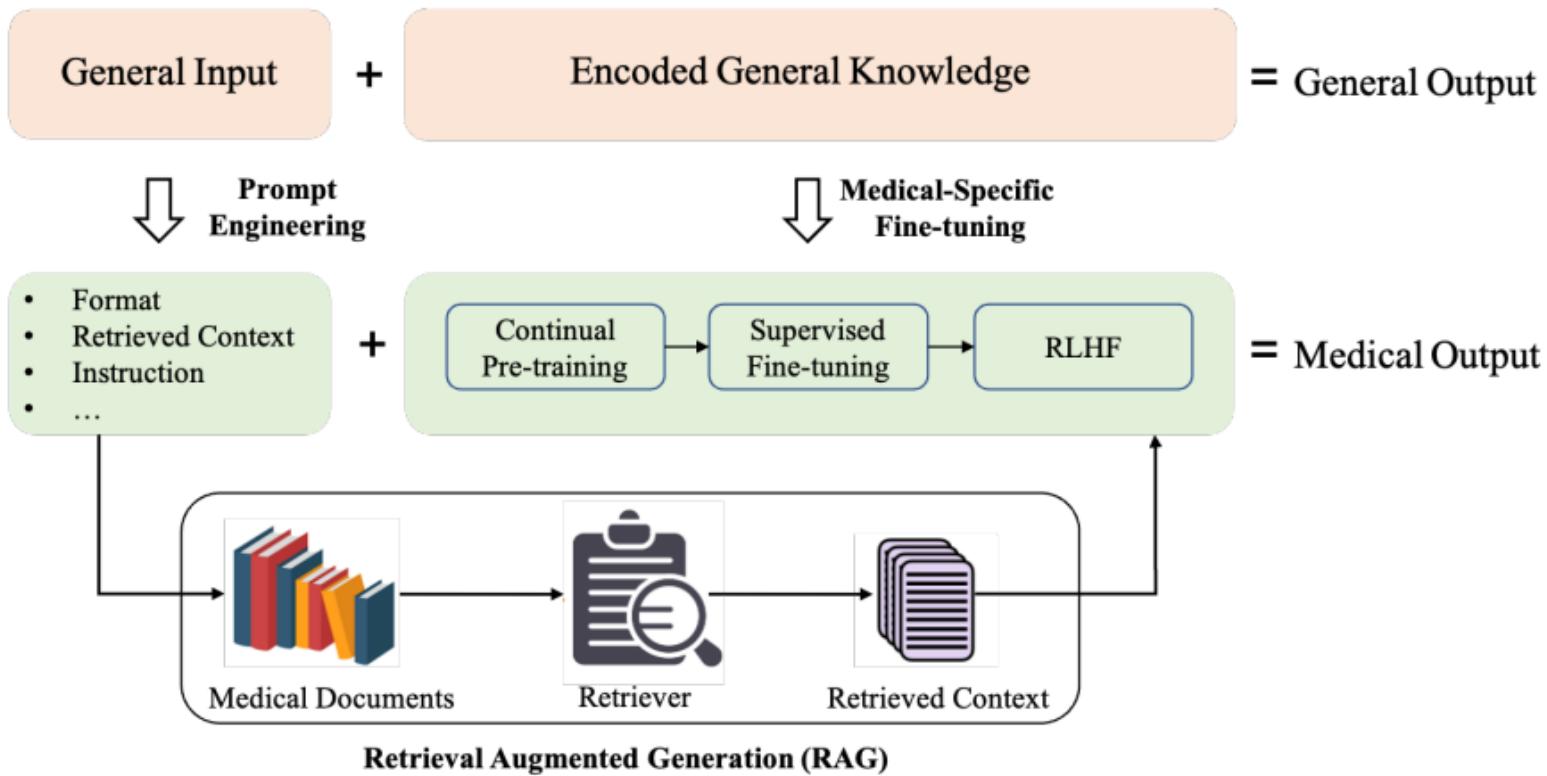


From General to Medical-specific LLMs



From General to Medical-specific LLMs

- Making a LLM to be a doctor

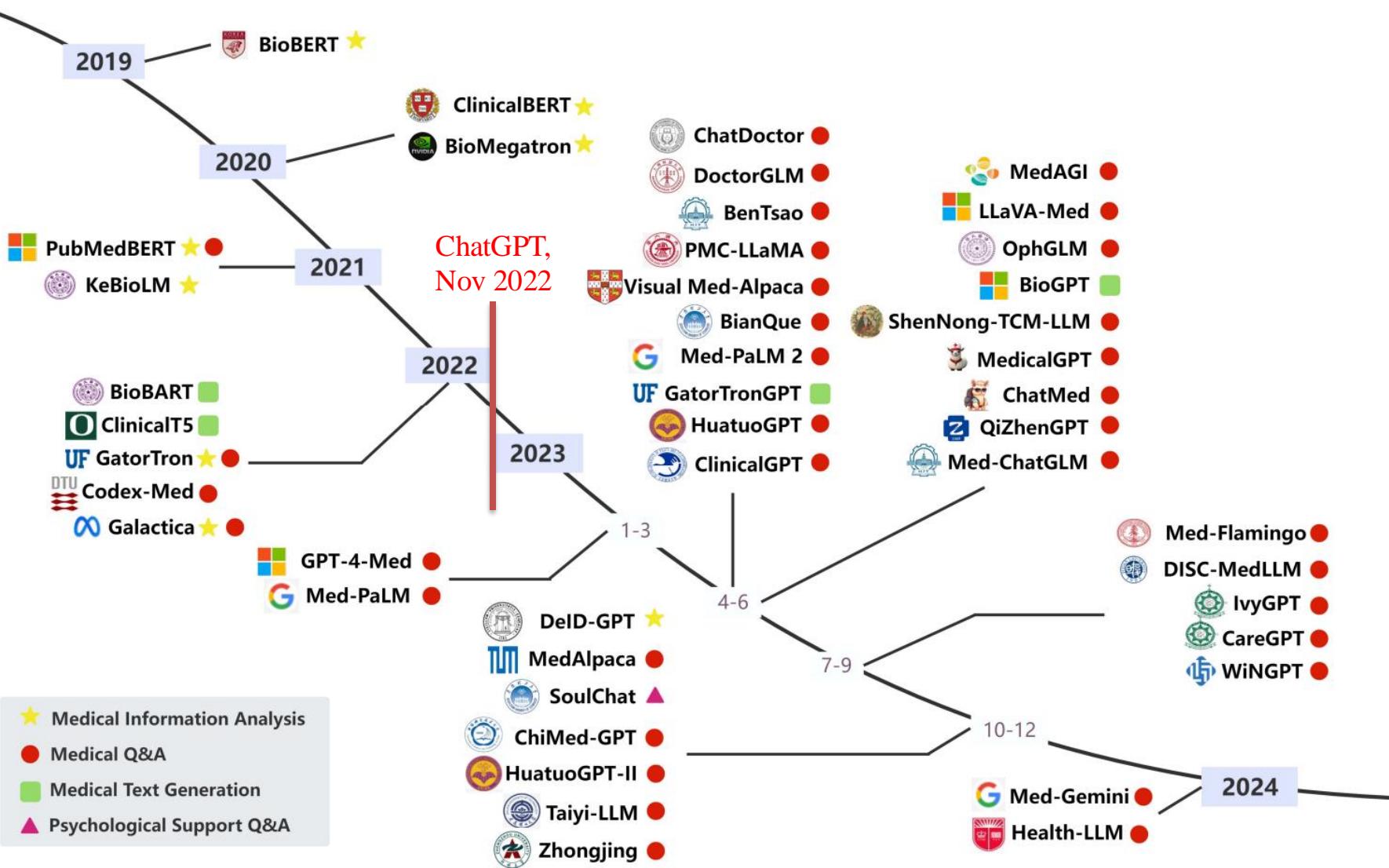


From General to Medical-specific LLMs

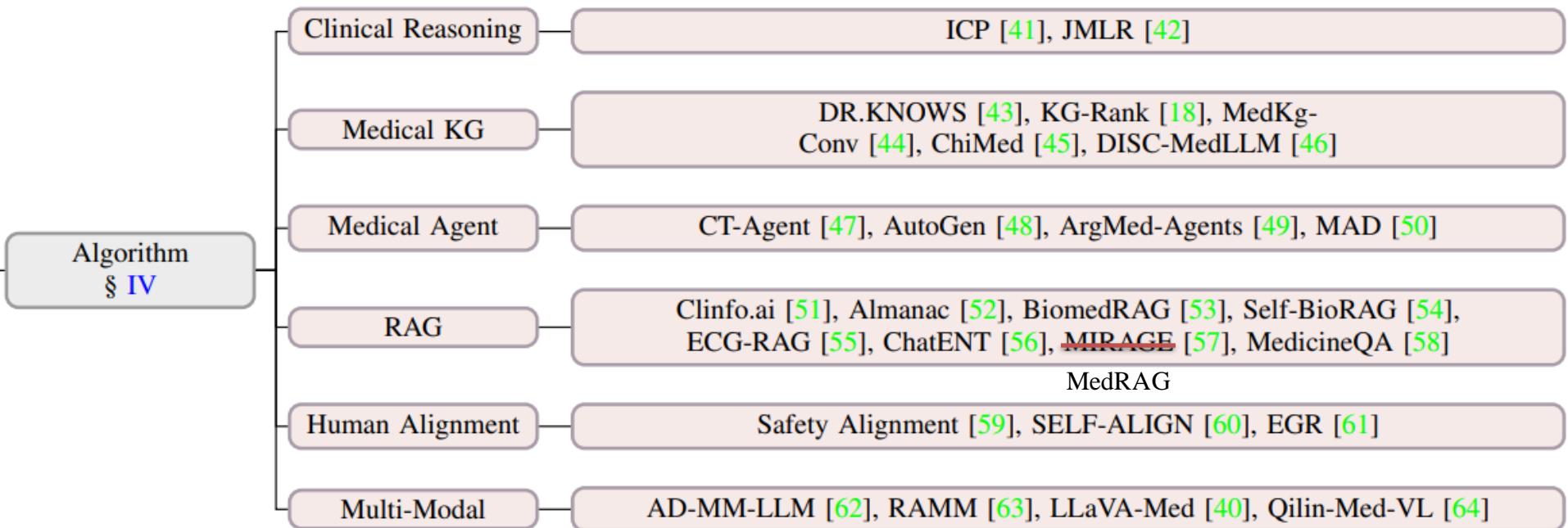
- Specific Med-LLMs

Model	Year	Method	Training Data
BioBERT [143]	2019	PT	Medical Abstracts and Articles
ClinicalBERT [144]	2020	PT	Clinical notes
BioMegatron [145]	2020	PT	Medical Literature
PubMedBERT [36]	2021	PT	Medical Literature
KeBioLM [146]	2021	PT	Medical Literature
.....			
WiNGPT [169]	2023	PT, IFT	Medical and General Knowledge
Taiyi-LLM [170]	2023	PT, IFT	Biomedical Datasets
Zhongjing [171]	2023	PT, SFT, RLHF	Medical Books, ChatMed, Medical Wiki
Med-Gemini [172]	2024	IFT	Medical Knowledge, Clinical Cases
Health-LLM [173]	2024	IFT	PMDATA, LifeSnaps, GLOBEM, AW_FB

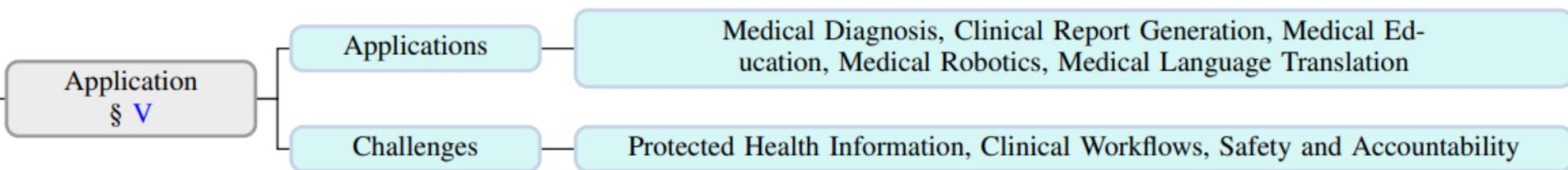
From General to Medical-specific LLMs



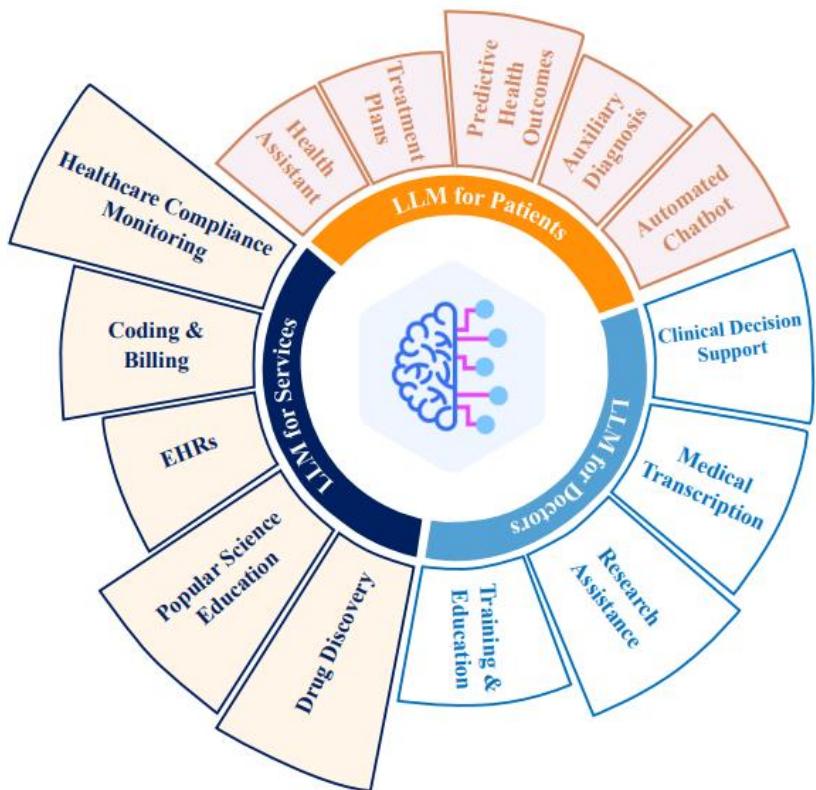
Improving Algorithms for Med-LLMs



Applying Medical LLMs



- Multifarious Applications



- Unique Challenges for Med-

- LLMs

- Protected Health Information
- Clinical Workflows
- Safety and Accountability

Trustworthiness and Safety

- Fairness
 - Research has revealed **biases in healthcare**, and LLMs are able to capture these biases from training data.
- Accountability
 - Lack of accountability in LLMs is recognized as an **obstacle** hindering its **application** in the medical field.
- Privacy
 - **Data privacy** is an important challenge for medical applications.
- Robustness
 - One future research direction is to explore the construction of effective **adversarial test samples** in the medical field to **evaluate the robustness** of large language models in the medical field.

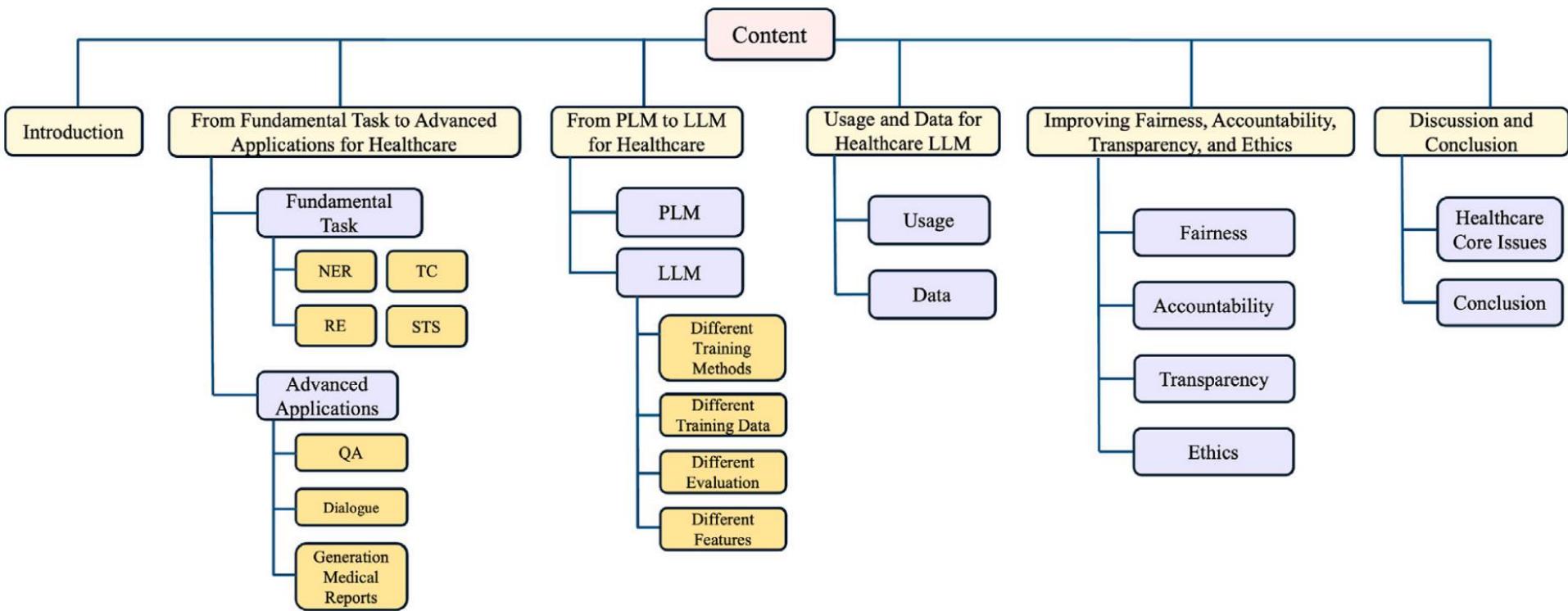
Future Directions

- Algorithmic advancements
 - E.g., multimodal learning, robot-assisted learning
- Industrial transformations
 - E.g., preventative and precision medicine, medical documentation
- Policy developments
 - Clear guidelines and standards for the *development, validation, and monitoring* of Med-LLMs

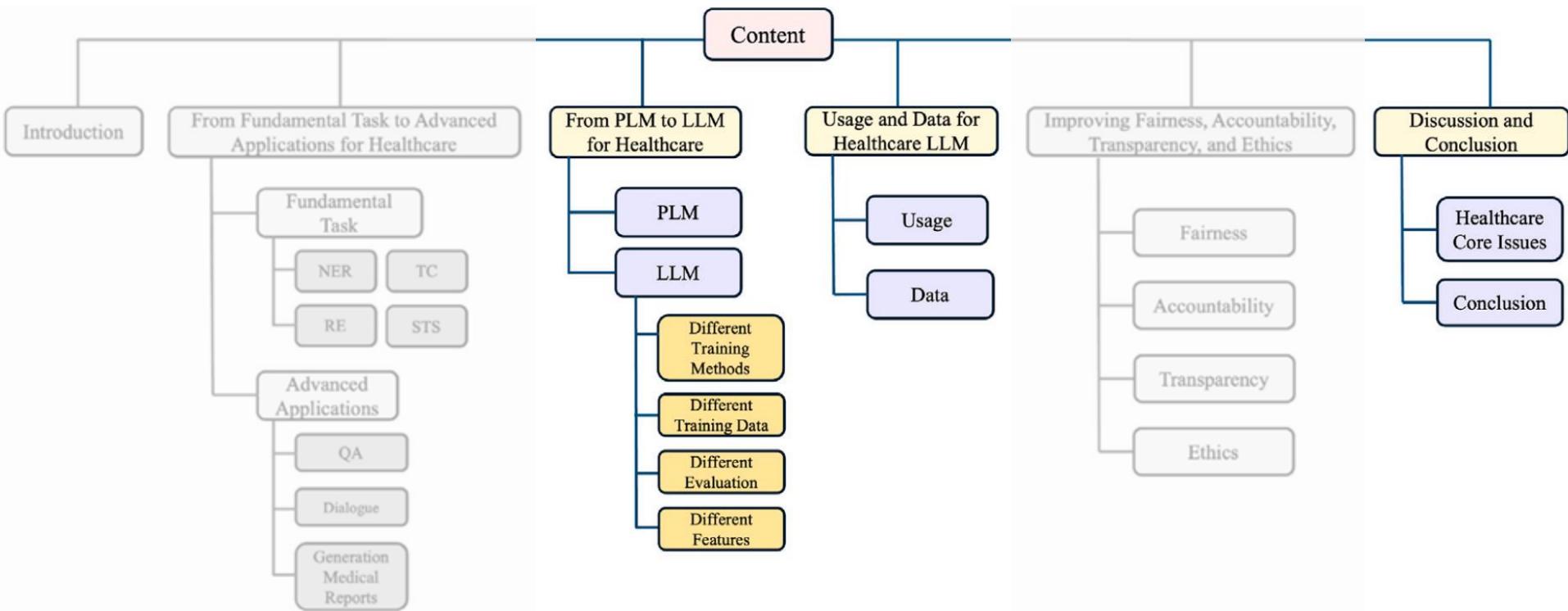
A Survey of Large Language Models for Healthcare: from Data, Technology, and Applications to Accountability and Ethics

HE ET AL.
NUS & NTU
JAN 2025

Organizational Framework



Organizational Framework



From PLMs to LLMs for healthcare

Summarization about the strengths and weaknesses of PLMs and LLMs by different tasks.

Task	PLMs features
Information extraction	Need labeled data
Text classification	Easy to adapt
Semantic textual similarity	Skilled at short contexts and fundamental tasks
Question answering	Limited language understanding and generation abilities
Dialogue system	Consist of multiple components
Report generation	Limited generation abilities and only single modality

LLMs features	Comparison
Zero-/few-shot	Have their own unique strengths
Explainable and reliable	LLMs have a slight advantage
Skilled at long contexts and complex tasks	Depend on text length
Better inherent professional knowledge	LLMs have a significant advantage
End-to-end system	LLMs have a significant advantage
Multimodal LLMs	LLMs have a significant advantage

From PLMs to LLMs for healthcare

- Summary of the performance for the three most popular datasets used to evaluate Healthcare LLMs

(%)	USMLE	MedMCQA	PubMedQA
FT-BERT	44.62 [67]	43.03 [60]	72.20 [67]
Galactica	44.60	77.60	77.60
PMC-LLaMA	44.70	50.54	69.50
GatorTronGPT	42.90	45.10	77.60
DoctorGLM	67.60	—	—
MedAlpaca	60.20	—	—
Codex	60.20	62.70	78.20
Med-PaLM	67.60	57.60	79.00
Med-PaLM	67.60	57.60	79.00
Aloe-Alpha	71.01	64.47	80.20
Med-PaLM 2	86.50	72.30	81.80
GPT-4	86.70	73.66	80.40
Human	87.00	90.00	78.00

Usage and data for healthcare LLM

- Usage
 - From Fine-tuning to In-context Learning
 - *From Medprompt to o1: Exploration of Run-Time Strategies for Medical Challenge Problems and Beyond*
 - From System 1 To System 2 – Chain-of-Thought
 - *From Medprompt to o1: Exploration of Run-Time Strategies for Medical Challenge Problems and Beyond*
 - AI Agents
 - *The rise of agentic AI teammates in medicine*

Usage and data for healthcare LLM

- Healthcare training data
 - EHR
 - E.g., MIMIC III, MIMIC IV, CPRD
 - Scientific Literature
 - E.g., PubMed, PubMed Central
 - Web Data
 - E.g., COMETA (from Reddit), WebText

Usage and data for healthcare LLM

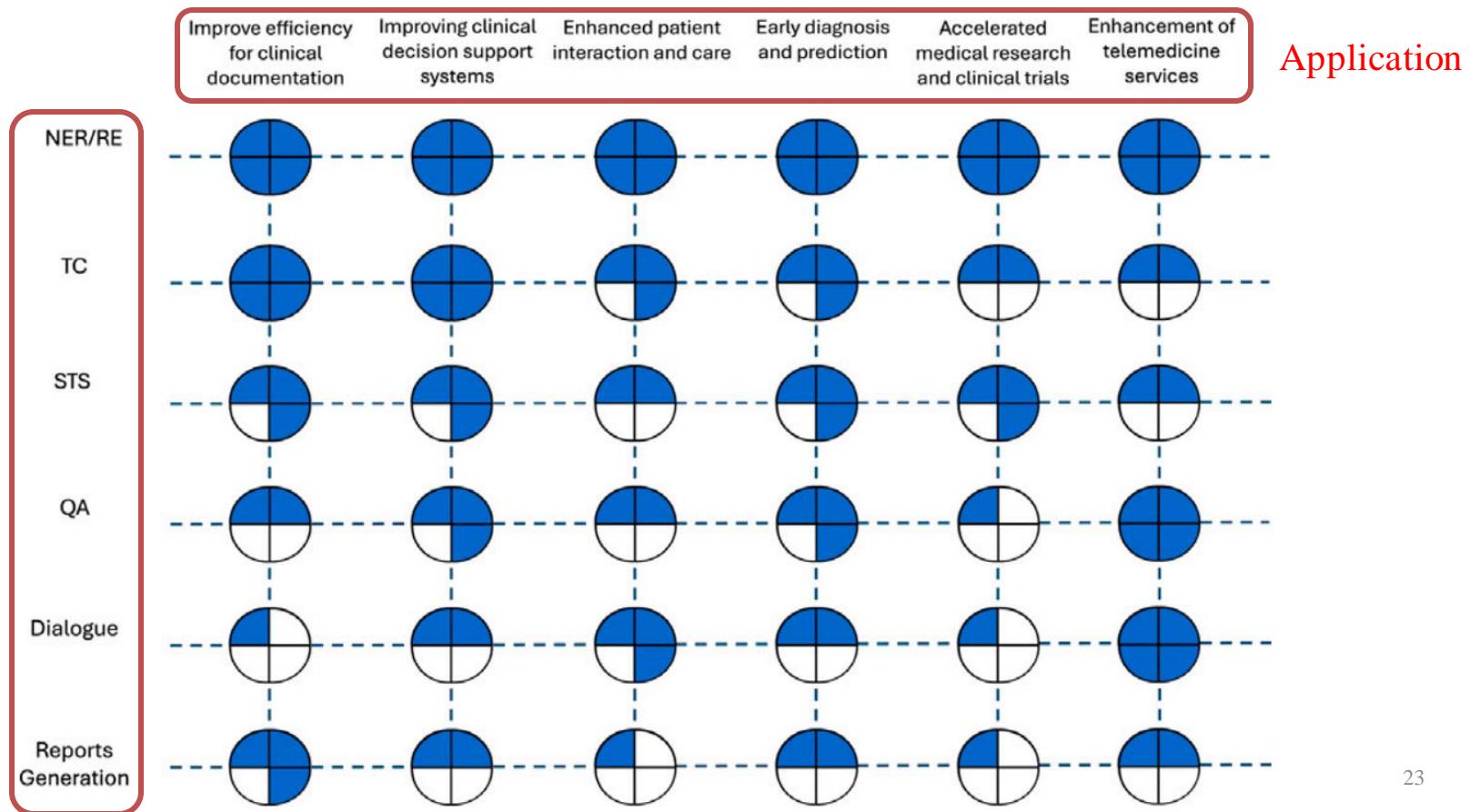
- Assessment of computation requirements

The statistics of computation cost for existing Healthcare LLM.

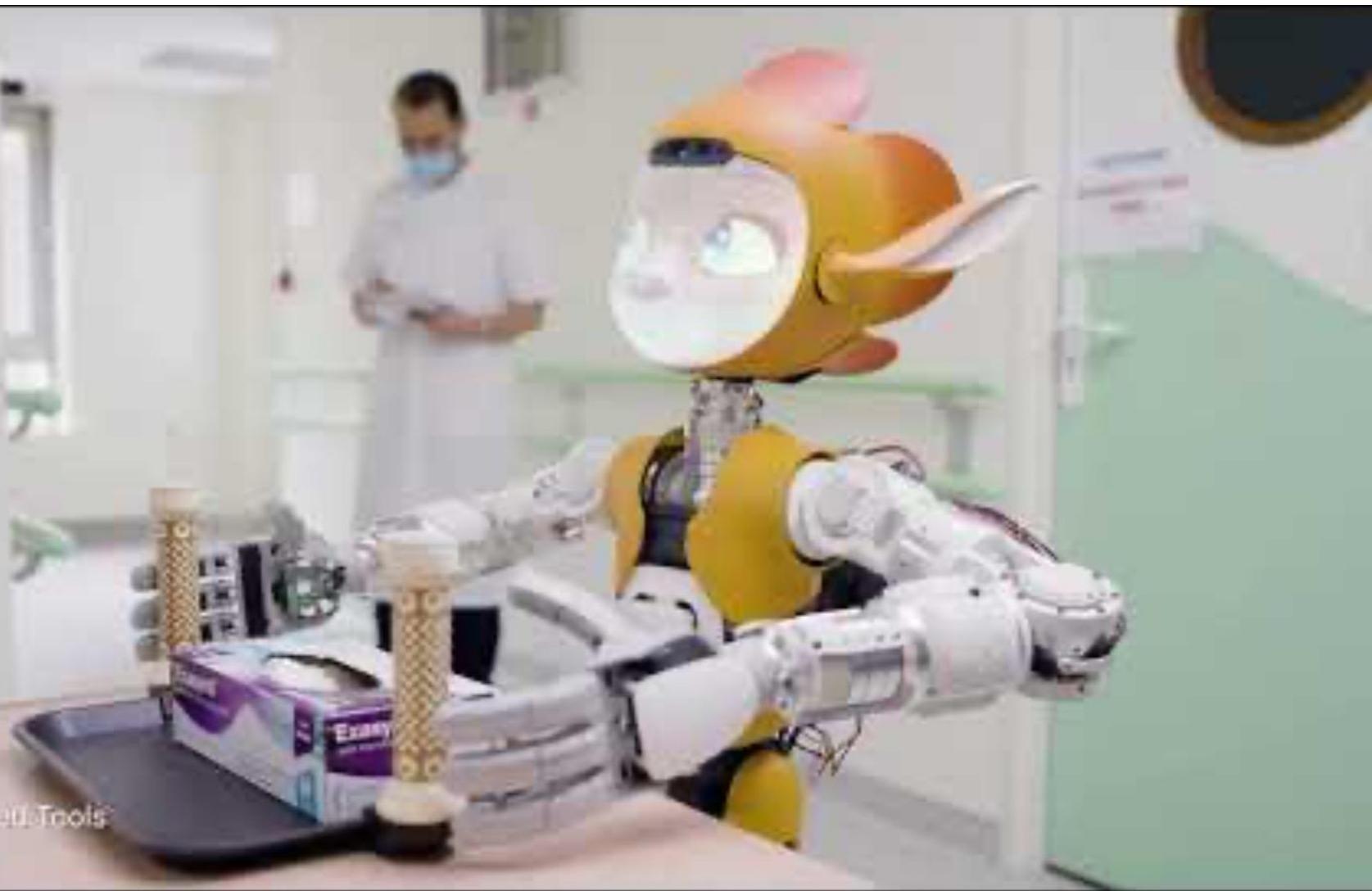
Model Name	Total data size	GPU type	GPU no.	GPU time
Visual Med-Alpaca	54k data points	A100-80G	4	2.51 h
GatorTron	>90 billion words	A100	992	6 days
Galactica	–	A100-80G	128	–
ChatDoctor	100k conversations	A100	6	3 h
DoctorGLM	3.5G	A100-80G	1	8 h
PMC-LLaMA	75B tokens	A100	8	7 days
Visual Med-Alpaca	44.8MB* (without images)	A100-80G	4	2.51 h
BianQue 1.0	9 million samples	RTX 4090	8	16 days
GatorTronGPT	277B tokens	A100-80G	560	26 days
HuatuoGPT	226,042 instances	A100	8	–
LLaVA-Med	15M image-caption pairs	A100	8	15 h
Med-Flamingo	1.3M image-caption pairs	A100-80G	8	6.75 days

Discussion

- Healthcare core issues
 - NLP technologies and their related healthcare applications (issues)



Discussion



Enchanted Tools

The rise of agentic AI teammates in medicine

JAMES ZOU, ERIC J TOPOLB

THE LANCET, 2025

Introduction

- From Tools to Teammates:
 - Shift from using AI as passive, **reactive** tools to proactive, **agentic** teammates.
 - AI agents can **autonomously** monitor healthcare systems, retrieve data, and track long-term patient histories
- Key agentic capabilities of AI agents
 - LLMs have rapidly improved their **reasoning** and problem-solving abilities
 - LLMs have become more adept at interacting with the external environment and using **tools**
 - **Collaborations** between multiple AI agents with complementary expertise have expanded their capability to address open-ended challenges

Example of Agentic AI in Medicine

- MedAgents (Tang et al., 2024)

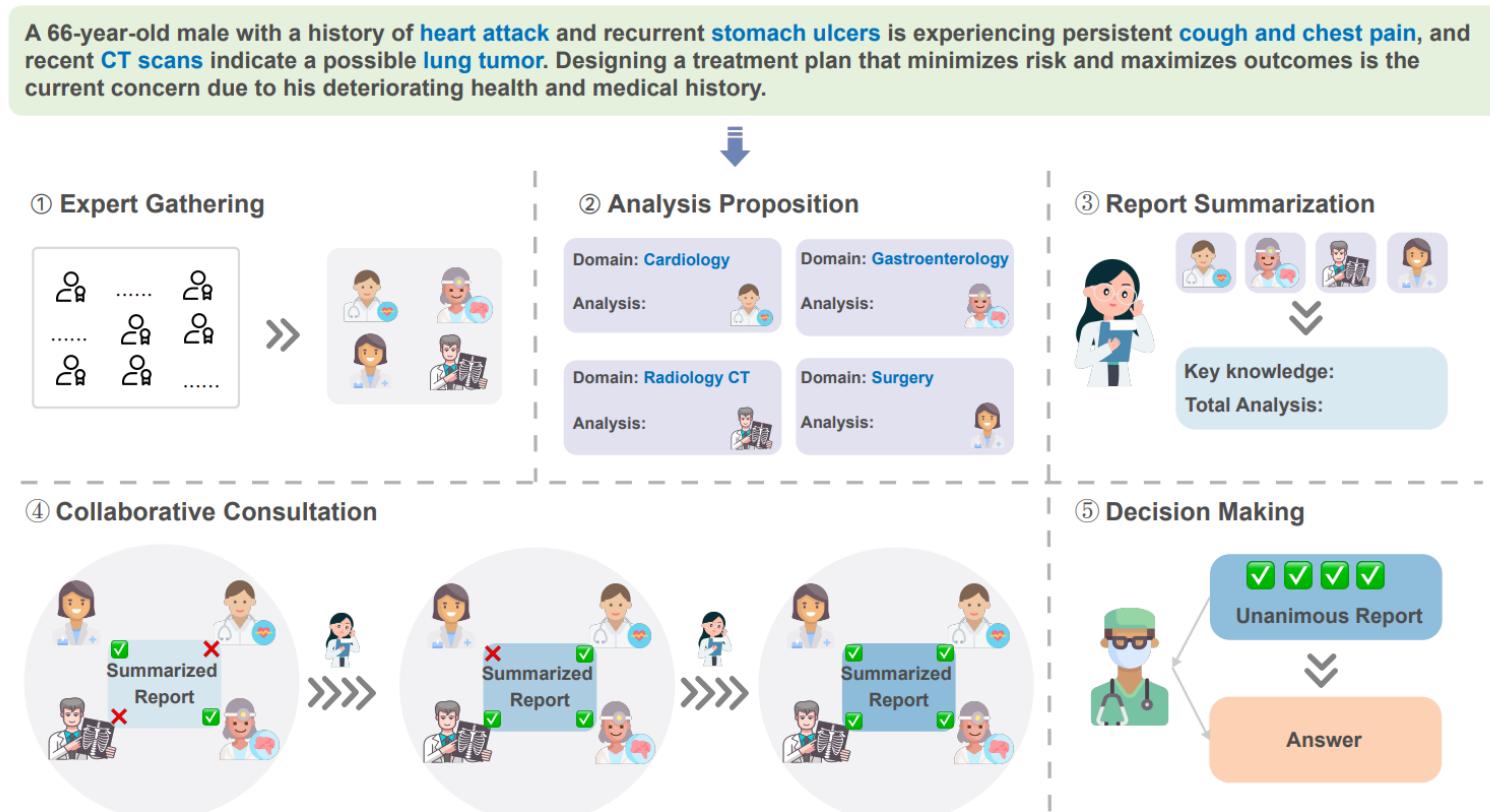


Figure 1: Diagram of our proposed MEDAGENTS framework. Given a medical question as input, the framework performs reasoning in five stages: (i) expert gathering, (ii) analysis proposition, (iii) report summarization, (iv) collaborative consultation, and (v) decision making.

Example of Agentic AI in Medicine

- MDAgents (Kim et al., 2024)

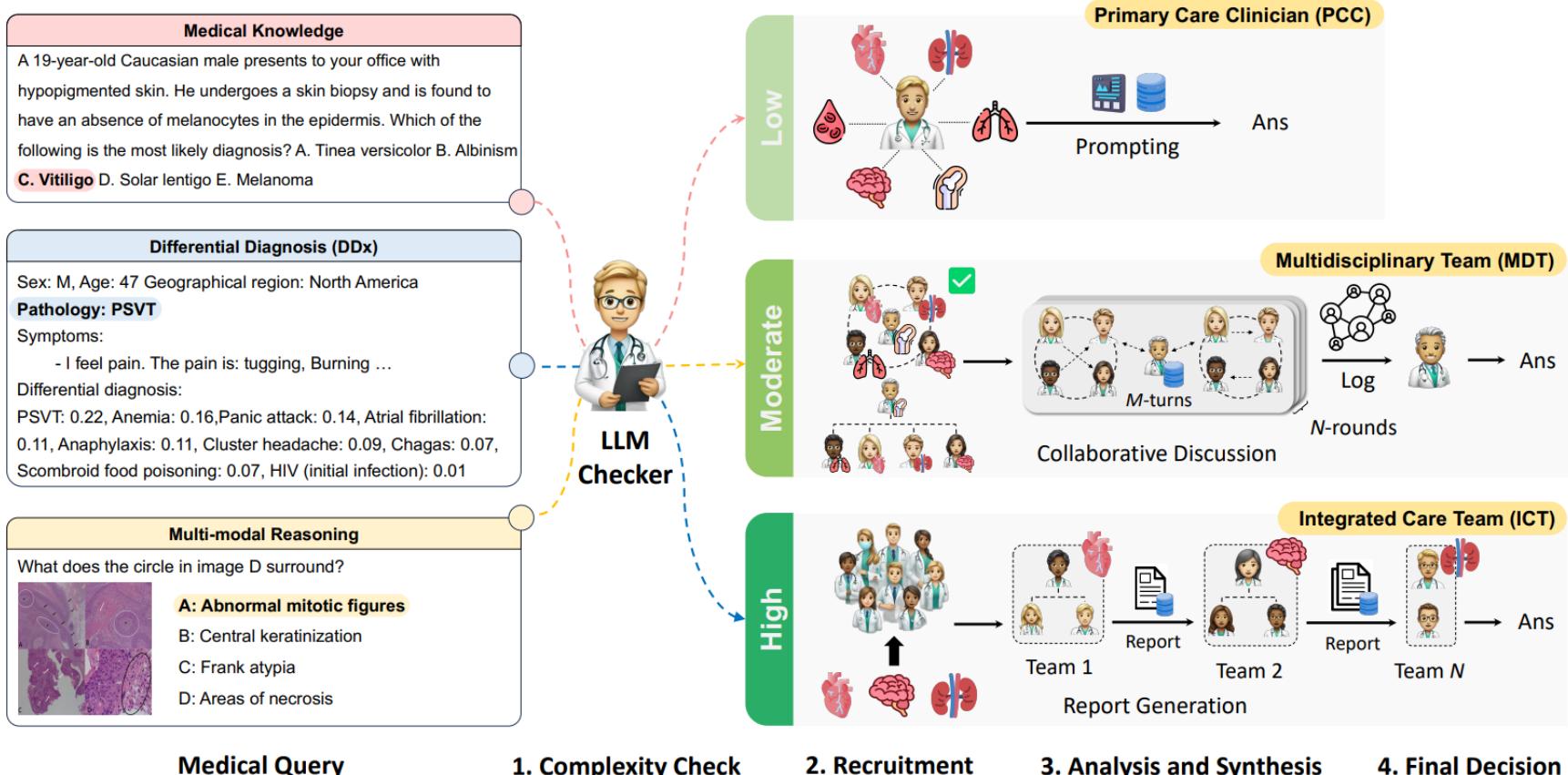
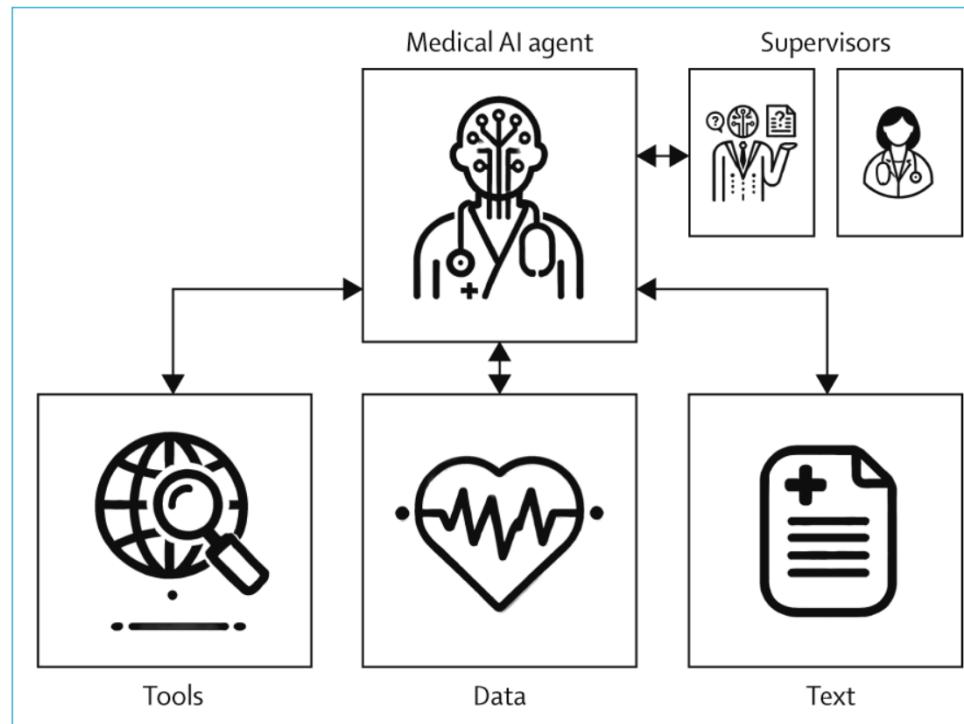


Figure 1: **Medical Decision-Making Agents (MDAgents) framework.** Given a medical query from different medical datasets, the framework performs 1) medical complexity check, 2) recruitment, 3) analysis and synthesis, and 4) decision-making steps.

Potentials of AI agents in Medicine

- Instead of juggling multiple tools (e.g., analyze medical images, search clinical guidelines), the clinician could interact with a single **manager agent**, which is able to orchestrate these tools.

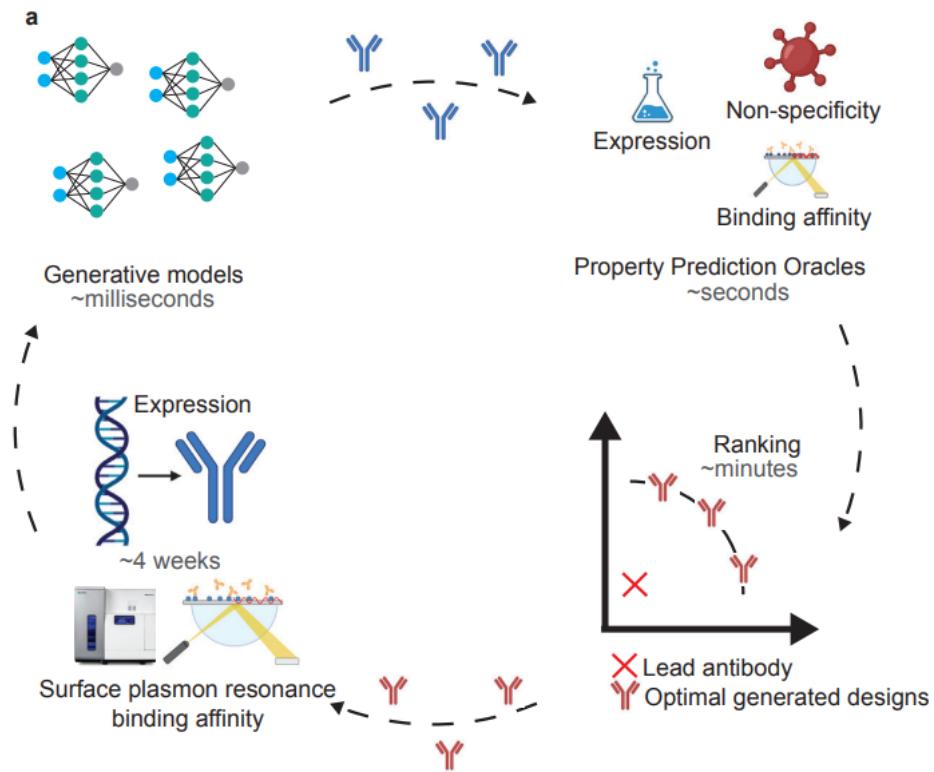


Safety and Regulation

- Frameworks for **evaluating** and regulating AI agents
 - Existing assessments typically focus on AI performance for a narrow medical output.
- **Confabulation** by AI agents
 - Continuous performance monitoring will be essential
- **Boundaries** for AI agent autonomy
 - Initial deployments might focus on low-risk administrative tasks in controlled virtual environments
- **Training** of healthcare professionals
 - Healthcare professionals would also need training to effectively collaborate with these semi-autonomous AI teammates.

More Example on GenAI for Medicine

- Problem:
 - Traditional **therapeutic antibody design** is expensive, time-consuming, and constrained by high-dimensional search spaces.
- Solution:
 - The paper introduces the **Lab-in-the-loop (LitL)** system, an AI-driven iterative optimization framework to automate and accelerate antibody design.



From Medprompt to o1: Exploration of Run-Time Strategies for Medical Challenge Problems and Beyond

NORI ET AL.
MICROSOFT & OPENAI
NOV 2024

Introduction

- Problem Statement:
 - Traditional **prompt engineering** methods (e.g., Medprompt) improve performance on medicine by using dynamic chain-of-thought (CoT) reasoning, few-shot prompting, and ensembling.
- New Paradigm:
 - The o1-preview model integrates run-time reasoning internally during training, making it “**reasoning-native**”.
 - This model **challenges the necessity of prompting strategies**.

Background: Medprompt

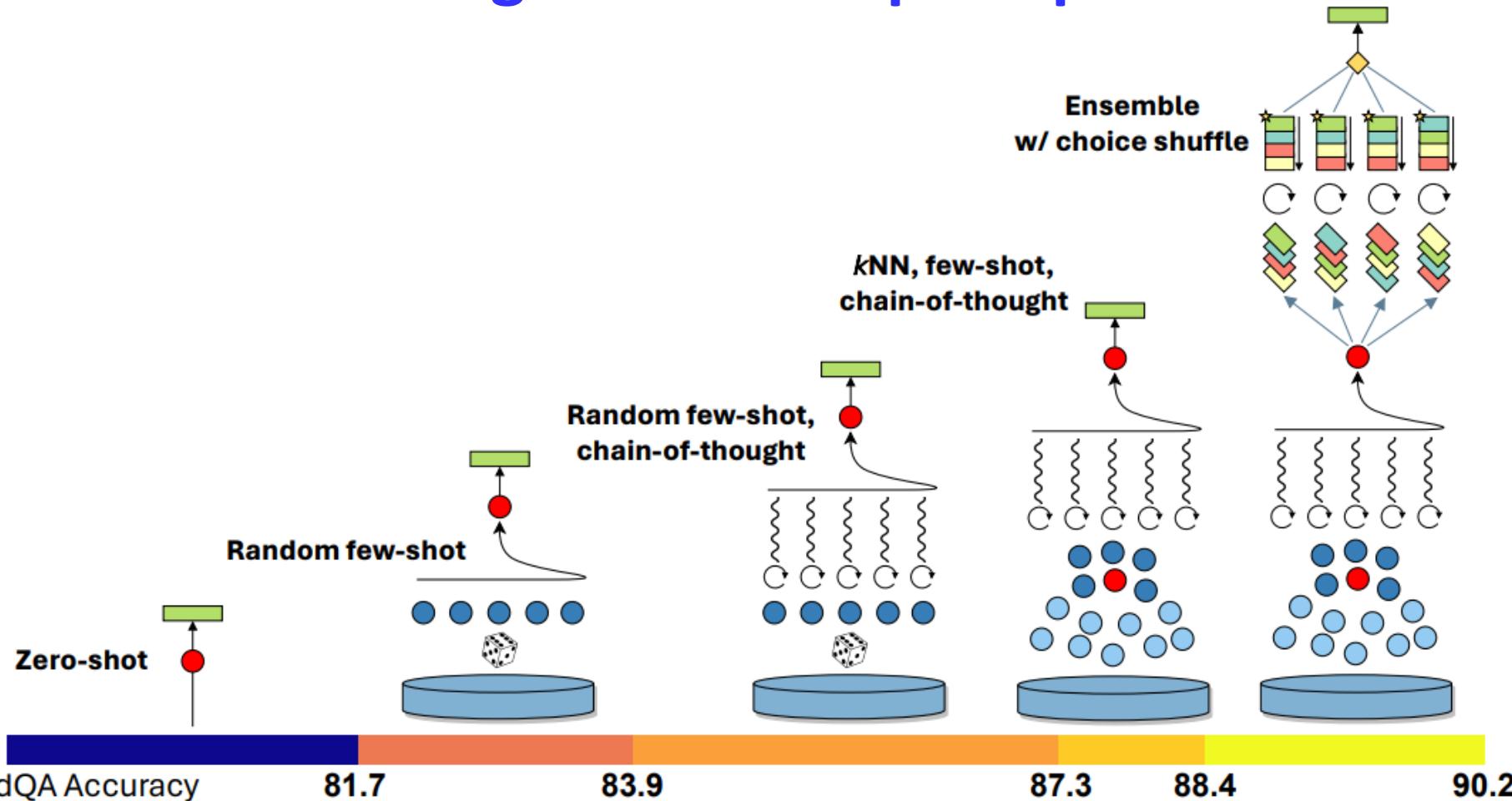


Figure 3: Visual illustration of Medprompt components and additive contributions to performance on MedQA. The prompting strategy combines $k\text{NN}$ -based few-shot example selection, GPT-4-generated chain-of-thought prompting, and answer-choice shuffled ensembling. Relative contributions of each component are shown at the bottom. Figure from [NLZ⁺23b]

Background: Medprompt

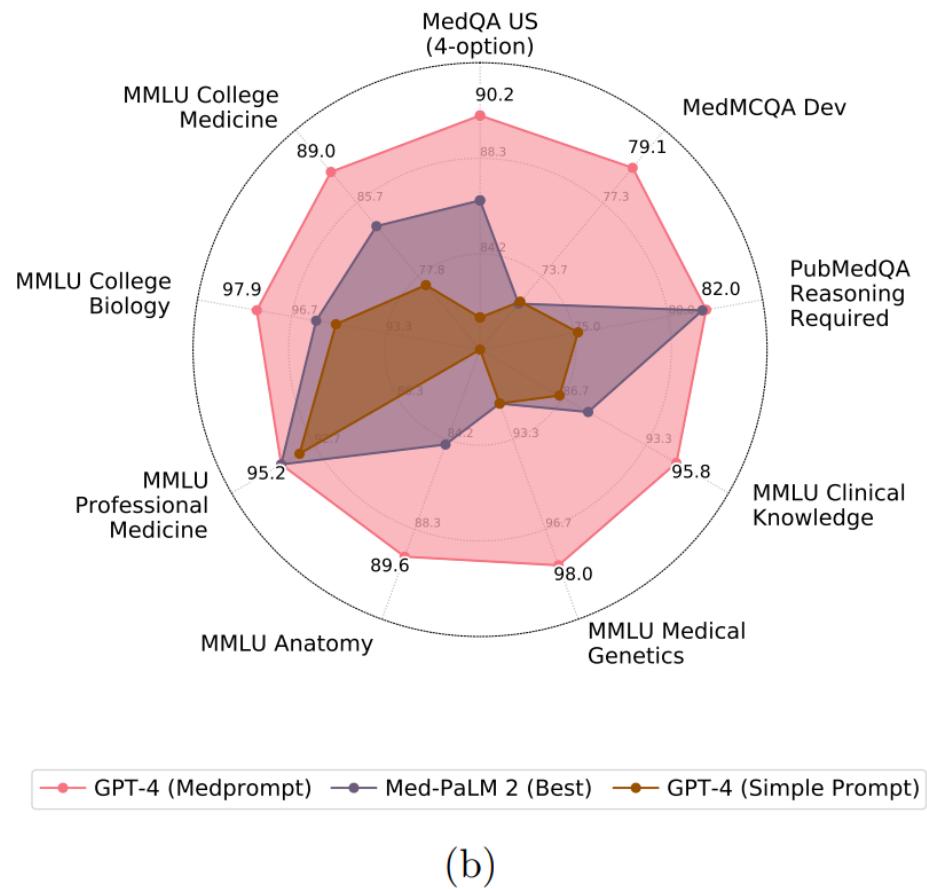
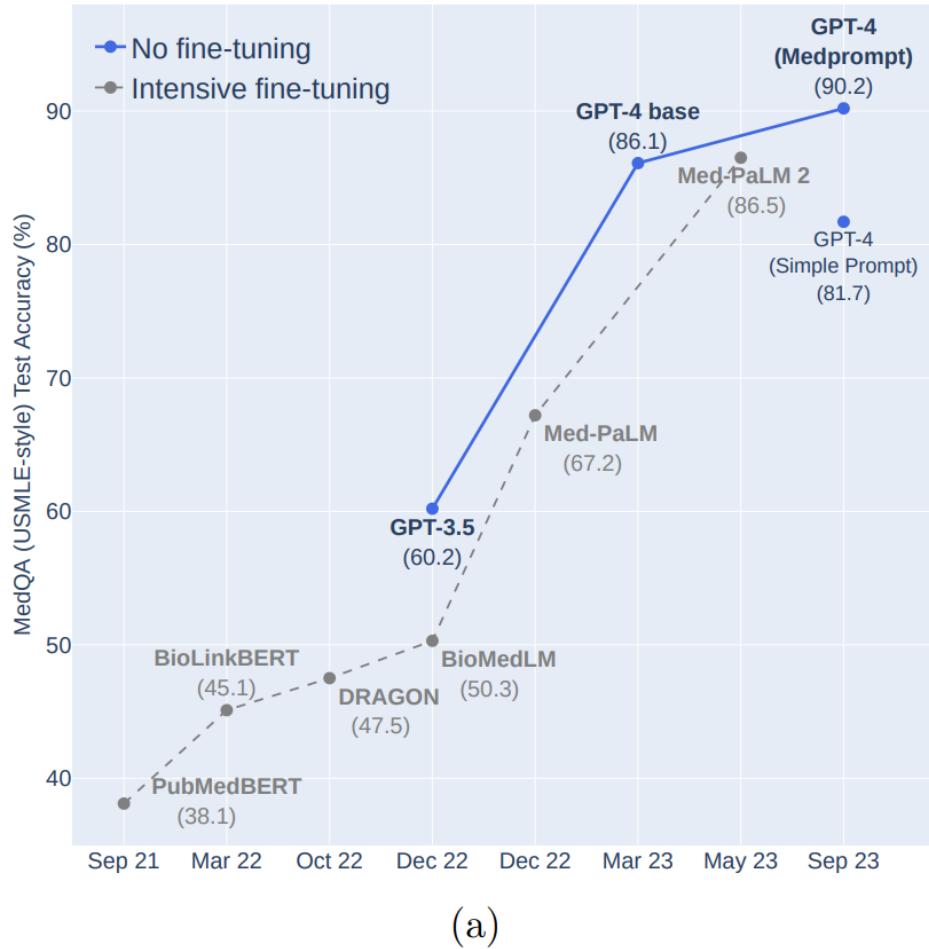


Figure 1: (a) Comparison of performance on MedQA. (b) GPT-4 with Medprompt achieves SoTA on a wide range of medical challenge questions.

Background: o1-preview

- Uses reinforcement learning to “think” before producing final responses.

Prompt

Suppose you are taking a medical exam. Please help me answer the following multiple-choice questions and give explanations.

“question”:“A child presented with abdominal distension, hepatomegaly, doll like facies and recurrent episodes of hypoglycemia. Which of the following is the most likely diagnosis?”

“opa”:“Menkes disease”,“opb”:“Down syndrome”,“opc”:“Von gierke disease”,“opd”:“Lesch nyhan syndrome”

O1-preview

To determine the most likely diagnosis for the child described, let's analyze the key clinical features provided:

1. **Abdominal distension and hepatomegaly:** These symptoms indicate an enlarged abdomen and liver, suggesting storage of substances in the liver.
2. **Doll-like facies:** This refers to a characteristic facial appearance with fat cheeks, often seen in some metabolic disorders.
3. **Recurrent episodes of hypoglycemia:** Frequent low blood sugar levels indicate a problem with glucose metabolism or regulation.

Now, let's examine each option:

Option A: Menkes Disease

- **Menkes disease** is an X-linked recessive disorder caused by defective copper absorption and transport due to mutations in the ATP7A gene.

• • • • •

*Evaluation of
OpenAI o1:
Opportunities and
Challenges of AGI*

Experimental Setup

- Medical benchmarks:
 - MedQA, MedMCQA, MMLU (Medical subset), NCLEX (Nurse licensing exam), JMLE-2024
- Official preparatory materials offered by NBME
 - USMLE Sample Exam
 - USMLE Self Assessment
- Evaluation metrics:
 - Accuracy for multiple-choice questions (MCQ)

Experimental Setup

- Setting: Quick Response

Quick Response

Please answer the following question as quickly as possible. We have narrowed down the possibilities to four different answers. I am in an emergency, and speed is of utmost importance. It is more important to answer quickly than it is to analyze too carefully.
Return just the answer as quickly as possible.

QUESTION

{question}

ANSWER CHOICES

{answer choices}

Please remember to answer quickly and succinctly. Time is of the essence!

Figure 10: Prompt that elicits the model to respond with less reasoning and completion tokens.

Experimental Setup

- Setting: Extended Reasoning

Extended Reasoning

Please answer the following multiple choice question. Take your time and think as carefully and methodically about the problem as you need to. I am not in a rush for the best answer; I would like you to spend as much time as you need studying the problem. When you're done, return only the answer.

QUESTION

{question}

ANSWER CHOICES

{answer choices}

Remember, think carefully and deliberately about the problem. Take as much time as you need. I will be very sad if you answer quickly and get it wrong.

Figure 11: Prompt that elicits the model to respond with more reasoning and completion tokens.

Experimental Setup

- Setting: Tailor Prompting

Tailored - Medical Questions

You are tasked with solving complex medical questions that assess both the knowledge and clinical reasoning required for a medical licensing exam. These questions cover critical topics such as anatomy, physiology, pathology, pharmacology, and patient management. Read the following question carefully and select the most accurate answer from the provided options.

****Question**:**

{question}

****Options**:**

{answer choices}

****Instructions**:**

- Think deeply and thoroughly, then choose the best possible answer from the given options (only one choice).
- Your final response must contain only the letter corresponding to the correct answer (e.g., "A"). Do not include explanations or additional text in your output.

****Answer**:**

Figure 12: Prompt for medical questions requiring knowledge and clinical reasoning.

Key Results & Findings

- Performance Gains:
 - o1-preview outperforms GPT-4 enhanced with Medprompt on several benchmarks.

Table 1: Performance of different models on a suite of medical benchmarks.

Dataset	GPT-4* (0 shot)	GPT-4-Turbo** (5 shot)	GPT-4o (0 shot)	GPT-4 Turbo** (Medprompt)	o1-preview (0 shot)
MedQA (US 4-option)	78.9%	81.4%	84.4%	90.2%	96.0%
JMLE-2024	-	87.3%	92.7%	92.7%	98.2%
MedMCQA Dev	69.5%	72.4%	76.4%	79.1%	83.9%
MMLU					
Clinical Knowledge	86.0%	86.4%	89.1%	95.8%	93.6%
Medical Genetics	91.0%	92.0%	96.0%	98.0%	99.0%
Anatomy	80.0%	80.0%	88.2%	89.6%	93.3%
Professional Medicine	93.0%	93.8%	97.4%	95.2%	97.0%
College Biology	95.1%	95.1%	95.1%	97.9%	98.6%
College Medicine	76.9%	76.9%	85.6%	89.0%	90.2%

* Results originally reported in [NKM⁺23] on the initial GPT-4 model release.

** Experimental results originally reported in [NLZ⁺23b], except JMLE-2024.

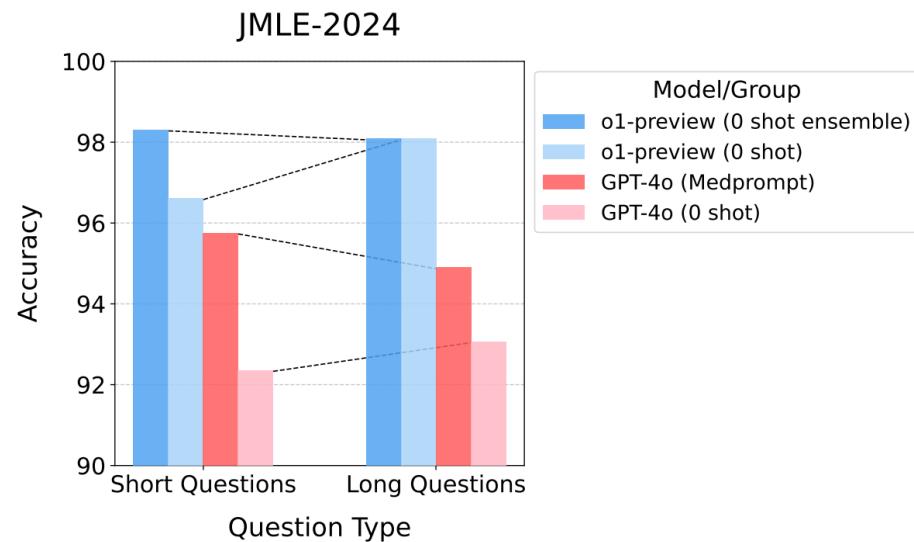
Key Results & Findings

- Performance Gains:
 - o1-preview outperforms GPT-4 enhanced with Medprompt on several benchmarks.

Table 2: Comparative analysis of performance of different models on USMLE Sample Exam and USMLE Self Assessment.

Dataset	GPT-4* (0 shot)	GPT-4o (0 shot)	o1-preview (0 shot)
USMLE Sample Exam			
Step 1	80.7%	89.8%	91.6%
Step 2	81.7%	89.1%	92.5%
Step 3	89.8%	93.1%	96.4%
USMLE Self Assessment			
Step 1	83.5%	87.5%	92.4%
Step 2	84.8%	91.8%	93.4%
Step 3	81.3%	80.6%	80.7%

* Results reported originally in [NKM⁺23] on the initial GPT-4 model release.



Key Results & Findings

- Impact of Prompting:
 - Five-shot prompting significantly decreases performance on MedQA

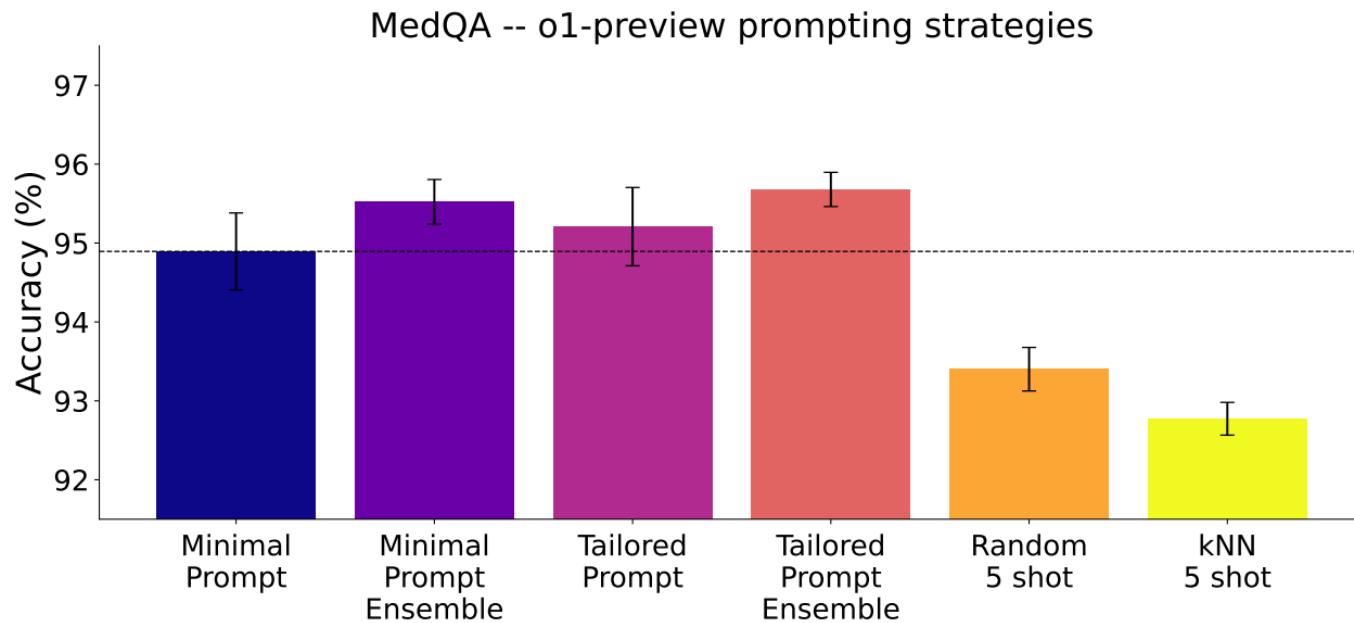
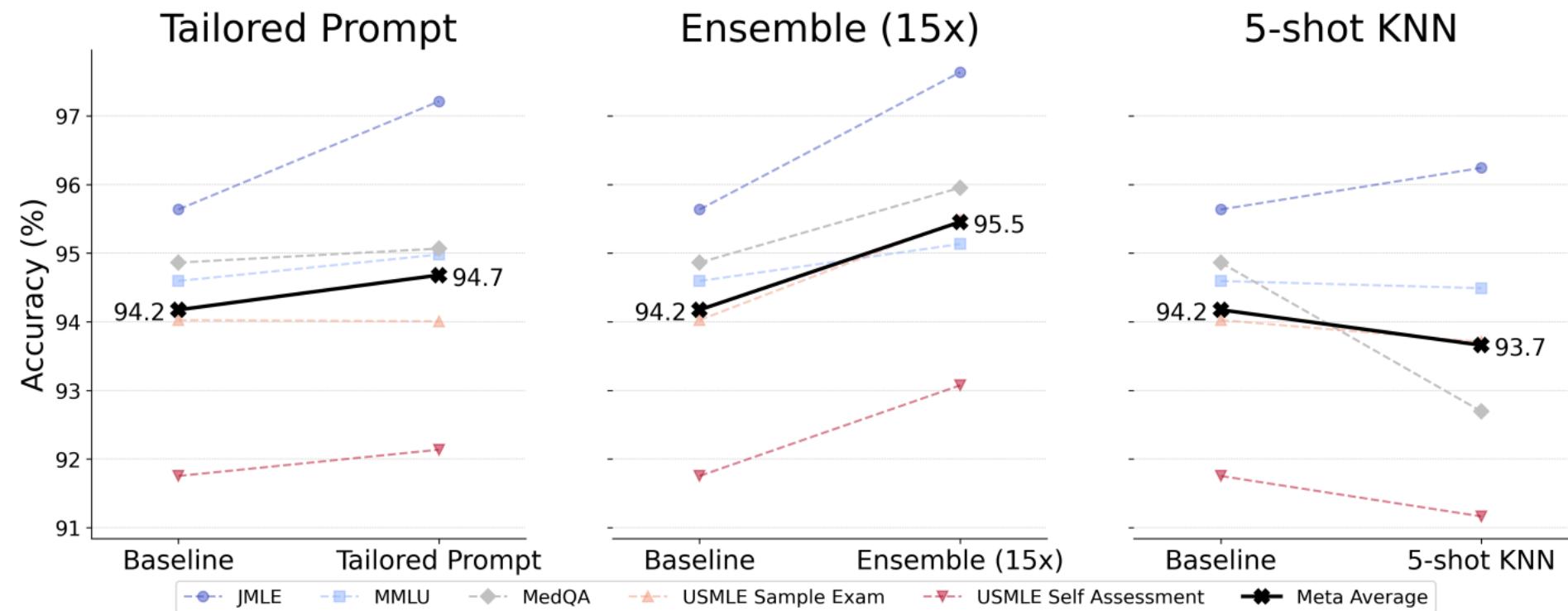


Figure 5: Comparison of prompting techniques on MedQA with the o1-preview model. Error bars indicate one standard deviation from three independent samples.

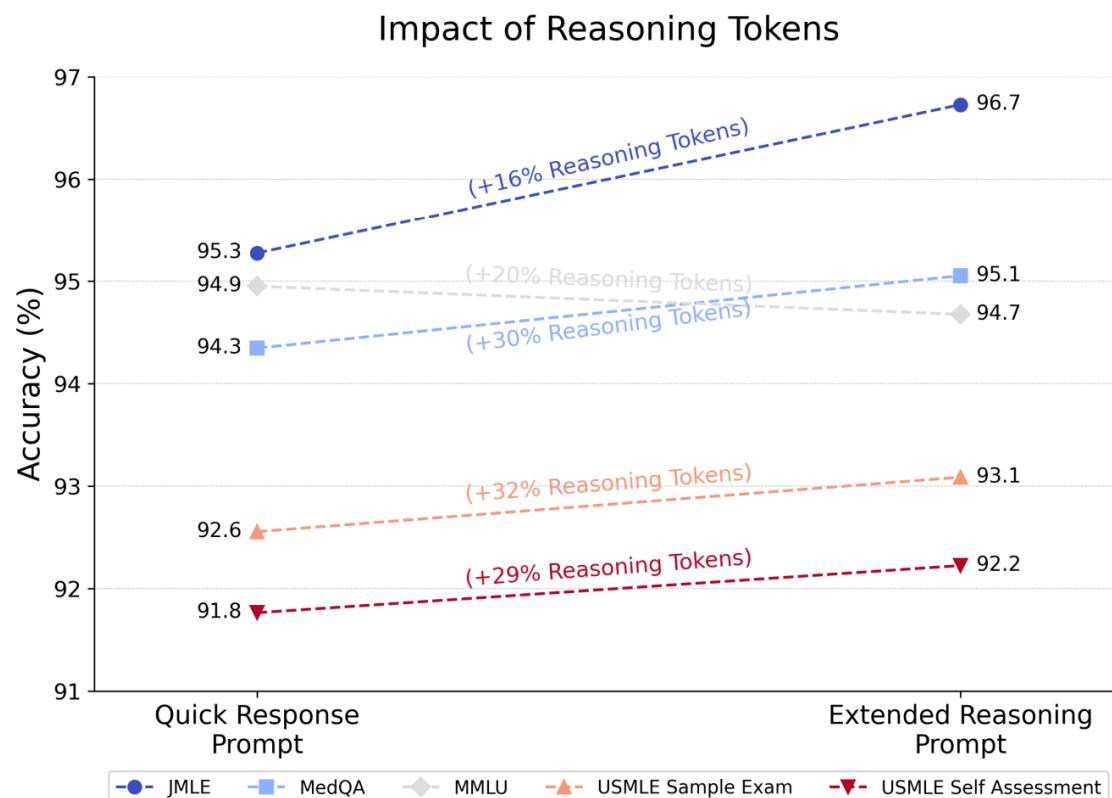
Key Results & Findings

- Impact of Prompting:
 - Ensemble improves performance across tasks



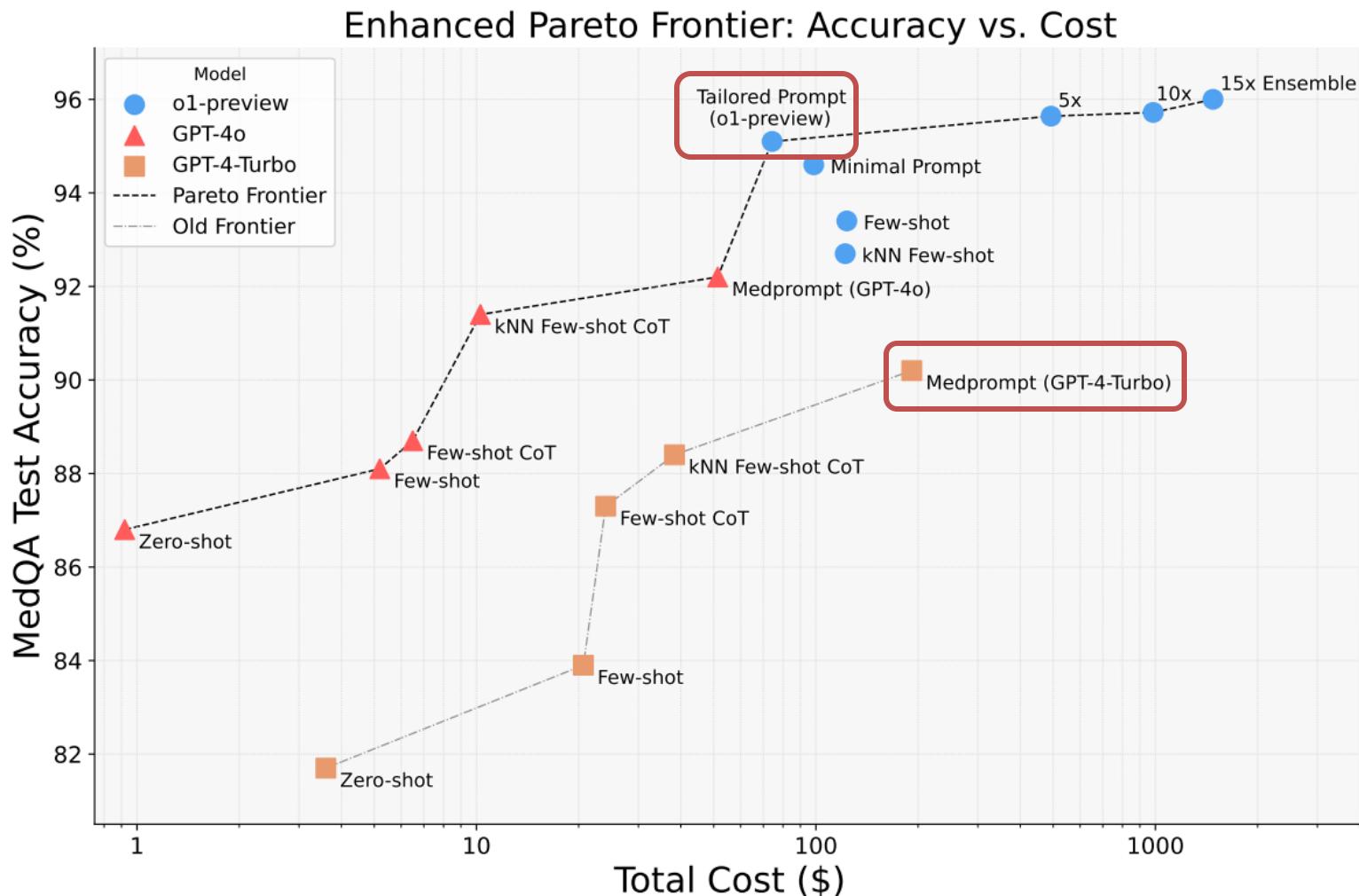
Key Results & Findings

- Role of Reasoning Tokens
 - Performance improves when the model is explicitly told to spend longer amounts of time reasoning



Key Results & Findings

- Accuracy and Cost Frontiers



Discussion & Future Directions

- Metareasoning Principles and Machinery
 - High-level metareasoning methods can facilitate runtime decision making by **dynamically allocating computing resources** across different generative processes and their combinations.
- Guiding LLM Inference and Sampling
 - Incorporating **token steering** mechanisms directly into model training may unlock further capabilities
 - E.g., entropy-based sampling techniques

Discussion & Future Directions

- Optimizing Input for LLMs
 - In-Context Learning
 - Determining how to efficiently provide **relevant examples** and **additional context** to optimize performance—especially in models such as o1-preview—remains a promising area of research
 - Integrating External Resources at Runtime
 - An essential avenue for further enhancing these models lies in their ability to actively acquire information at run-time from **external sources** such as the web and knowledge bases (KBs)

Discussion & Future Directions

- Reasoning
 - Chain-of-Thought prompting
 - Structured problem-solving
 - ReAct, skeleton-based prompting, and tree-based reasoning
 - Training LLMs to improve real-time reasoning
 - Self-Taught Reasoner (STaR)
 - Process supervision
 - Let's Verify Step-by-Step (Process reward model)
 - Scaling test-time computation
 - Reinforcement learning

Discussion & Future Directions

- Leveraging Multiple Runs and Models
 - Ensembling
 - Simple majority voting is a popular approach for aggregating outputs
 - Sophisticated methods are emerging
 - Ensemble Refinement, LLM-Blender
 - A major challenge with ensembling is the **computational cost**
 - Model Federation and Multi-Agent Architectures
 - Agent frameworks and multi-agent orchestration enable models to **dynamically** select and integrate the tools required to solve a given problem

References

- Zou, J., & Topol, E. J. (2025). The rise of agentic AI teammates in medicine. *The Lancet*, 405(10477), 457.
- Frey, N. C., Hotzel, I., Stanton, S. D., Kelly, R. L., Alberstein, R. G., Makowski, E. K., ... & Gligorijevic, V. (2025). Lab-in-the-loop therapeutic antibody design with deep learning. *bioRxiv*, 2025-02.
- Tang, X., Zou, A., Zhang, Z., Li, Z., Zhao, Y., Zhang, X., ... & Gerstein, M. (2024, August). MedAgents: Large Language Models as Collaborators for Zero-shot Medical Reasoning. In *Findings of the Association for Computational Linguistics ACL 2024* (pp. 599-621).
- Kim, Y., Park, C., Jeong, H., Chan, Y. S., Xu, X., McDuff, D., ... & Park, H. (2024). Mdagents: An adaptive collaboration of llms for medical decision-making. *Advances in Neural Information Processing Systems*, 37, 79410-79452.
- Nori, H., Usuyama, N., King, N., McKinney, S. M., Fernandes, X., Zhang, S., & Horvitz, E. (2024). From medprompt to o1: Exploration of run-time strategies for medical challenge problems and beyond. *arXiv preprint arXiv:2411.03590*.
- Zhong, T., Liu, Z., Pan, Y., Zhang, Y., Zhou, Y., Liang, S., ... & Liu, T. (2024). Evaluation of openai o1: Opportunities and challenges of agi. *arXiv preprint arXiv:2409.18486*.
- Nori, H., Lee, Y. T., Zhang, S., Carignan, D., Edgar, R., Fusi, N., ... & Horvitz, E. (2023). Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*.
- Liu, L., Yang, X., Lei, J., Liu, X., Shen, Y., Zhang, Z., ... & Ren, K. (2024). A survey on medical large language models: Technology, application, trustworthiness, and future directions. *arXiv preprint arXiv:2406.03712*.
- He, K., Mao, R., Lin, Q., Ruan, Y., Lan, X., Feng, M., & Cambria, E. (2025). A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *Information Fusion*, 102963.