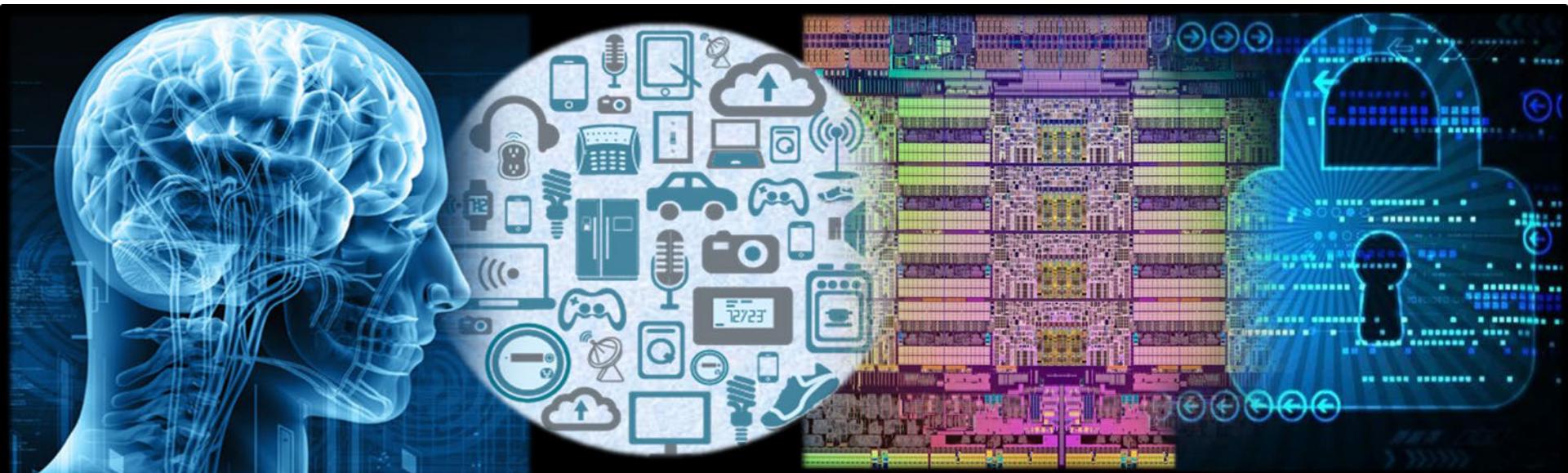


Applications of LLMs in Medical Domain

(Generative AI Workshop)

Abdul Basit

eBrain Lab, New York University (NYU) Abu Dhabi, UAE



History of Medicine

Ancient Greek

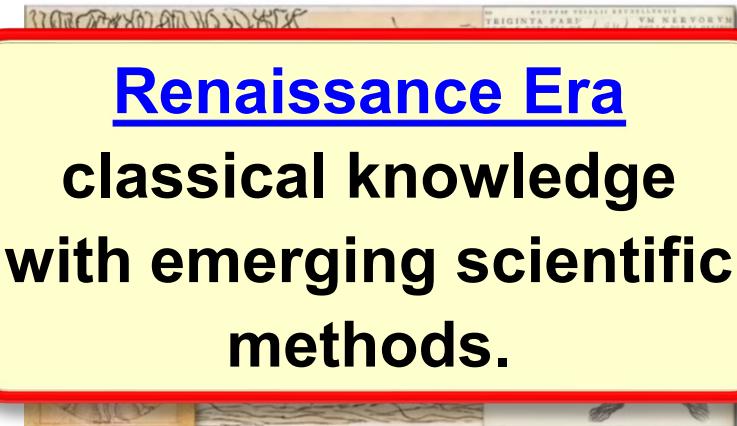
Primitive Technology and Methods etc.



Source: <https://www.tes.com/teaching-resource/ancient-greek-medicine-12326626>

Renaissance Era

classical knowledge with emerging scientific methods.



Source: <https://www.medicalnewstoday.com/articles/323533>

c. 1780

c. 1870

c. 1970

c. 2010

>= 2020

Medicine in Modern Era

Advanced technology and evidence-based practices revolutionizing healthcare and treatment.

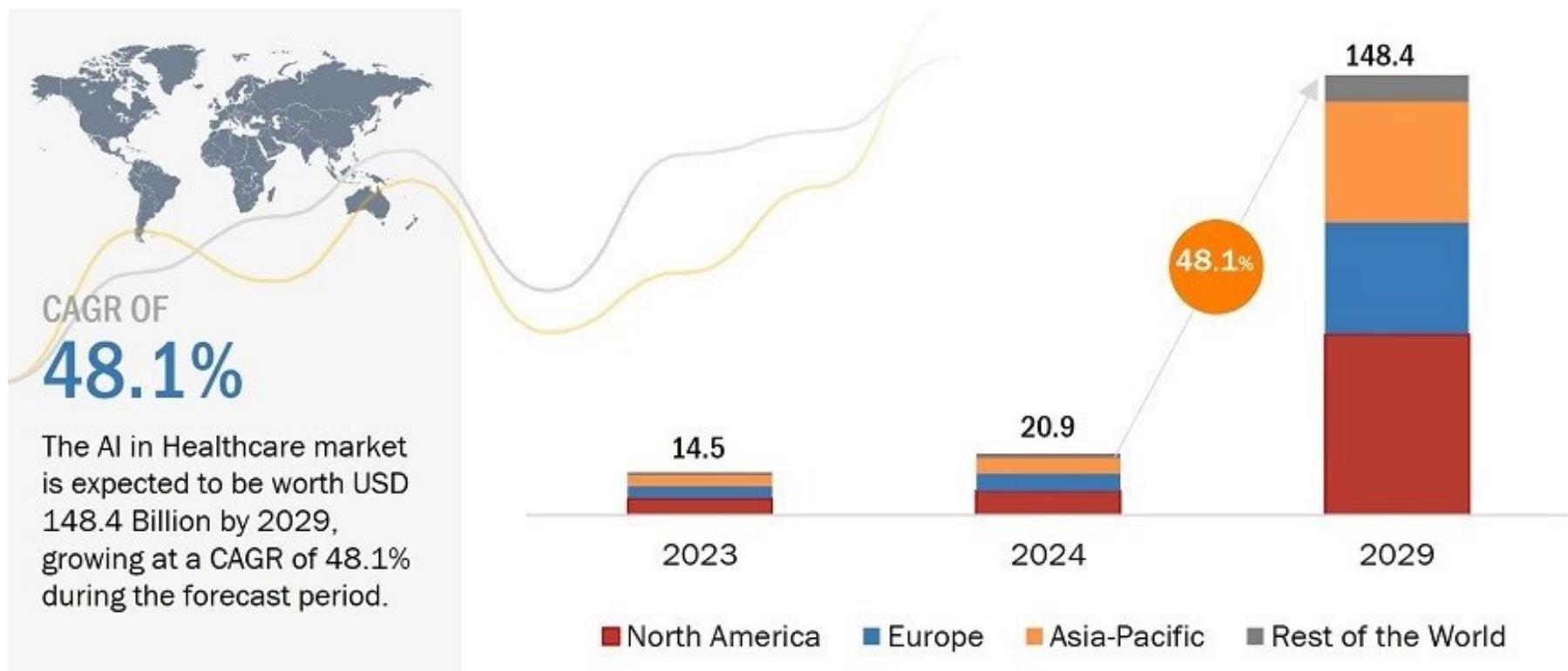
Healthcare 1.0
Vaccine
Sanitation
Germ Theory

Source: [Reliable and Resilient AI and IoT-based Personalised Healthcare Services: A Survey](https://www.semanticscience.org/perspectives/reliable-and-resilient-ai-and-iot-based-personalised-healthcare-services-a-survey)



AI in Healthcare Market (USD BN)

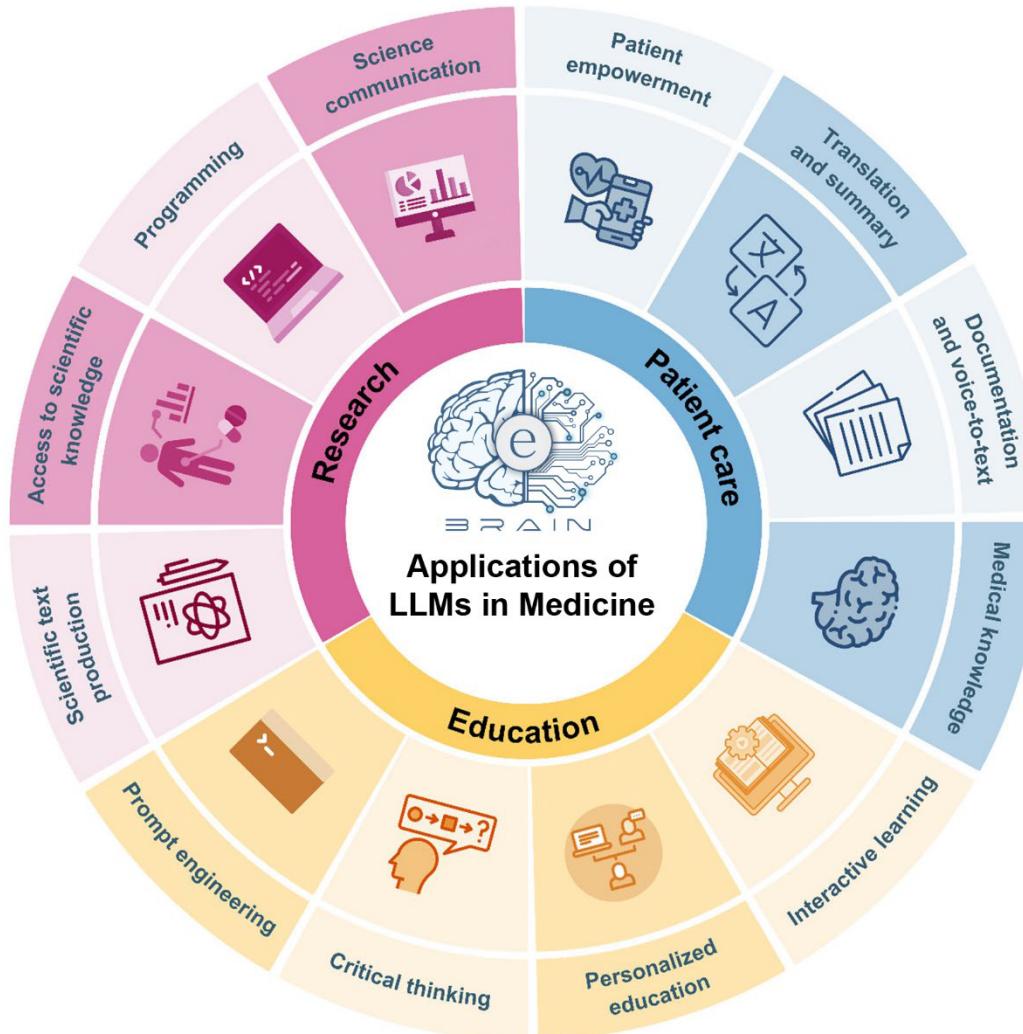
- The AI in Healthcare market is expected to grow from **USD 20.9 billion** in 2024 to **USD 148.4 billion** by 2029, with a CAGR of 48.1%, driven by the need for better healthcare services, enhanced computing power, and decreasing hardware costs.



Source <https://www.marketsandmarkets.com/Market-Reports/artificial-intelligence-healthcare-market-54679303.html>

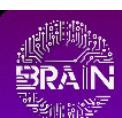
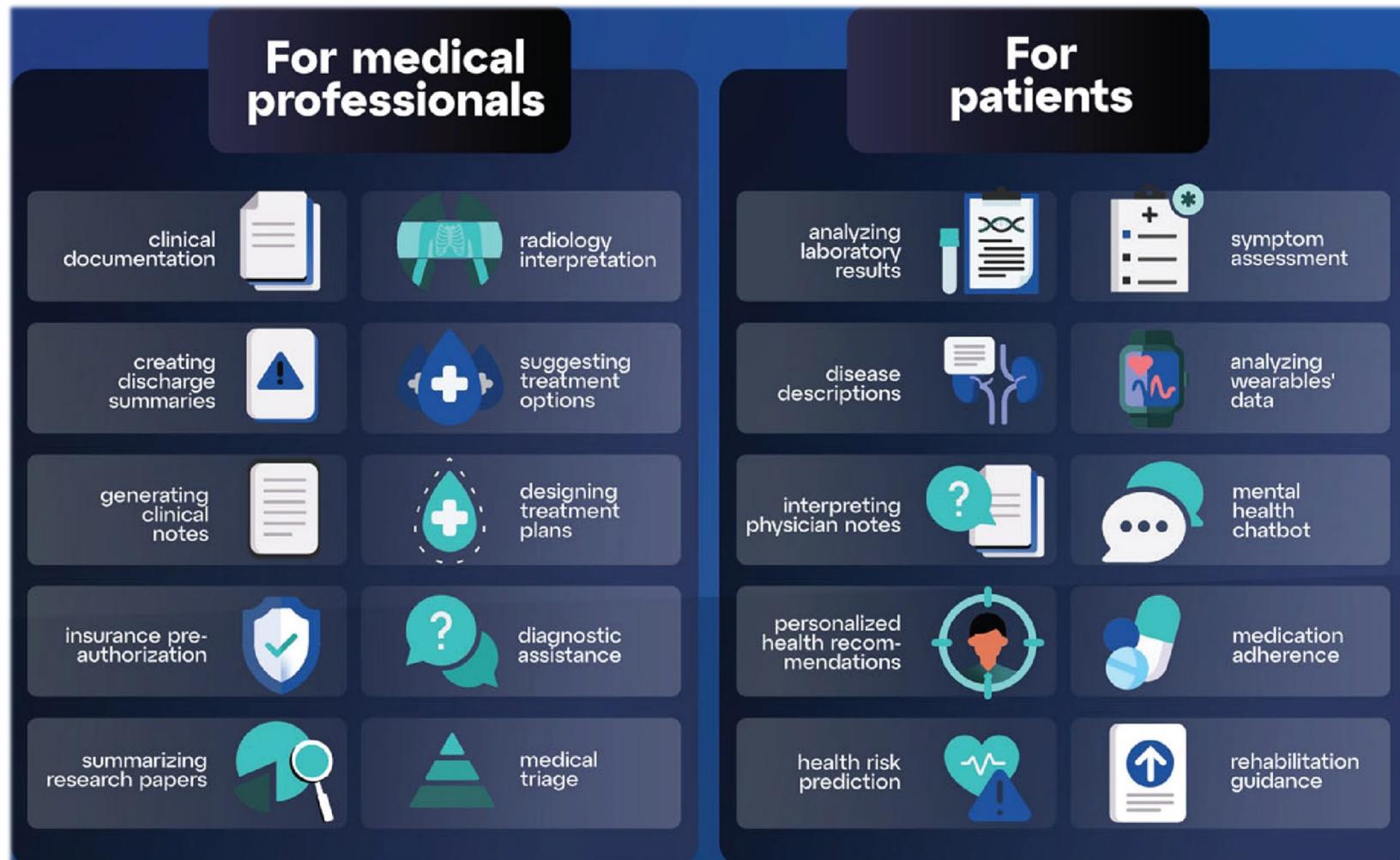
Introduction to LLMs in Healthcare

□ **LLMs in Medicine:** Revolutionizing medical informatics, diagnosis, patient care, and Education



Introduction to LLMs in Healthcare

- Enhancing care through diverse applications for both medical professionals and patients, leading to improved diagnostics and personalized treatment.



Introduction to LLMs in Healthcare

- Enhancing care through diverse applications for both medical professionals and patients, leading to improved diagnostics and personalized treatment.

User dependency and trust

Lack of physical examination

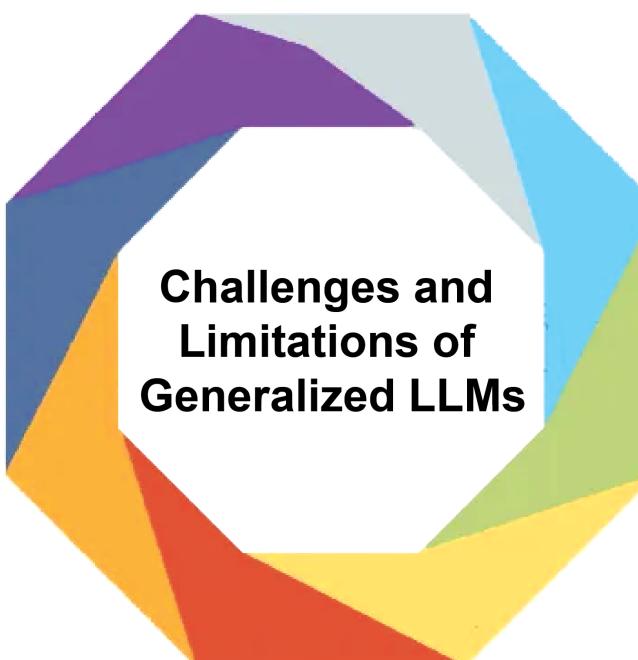
Language and cultural barriers

Limited real-time interaction

Lack of medical expertise

Hallucination

Bias and fairness

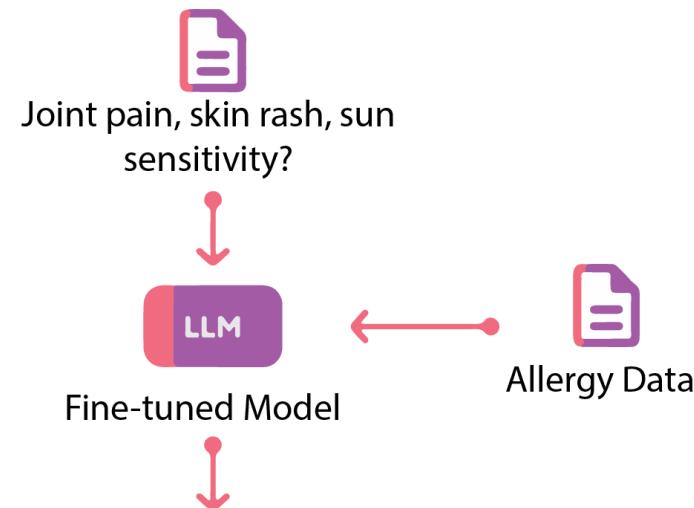
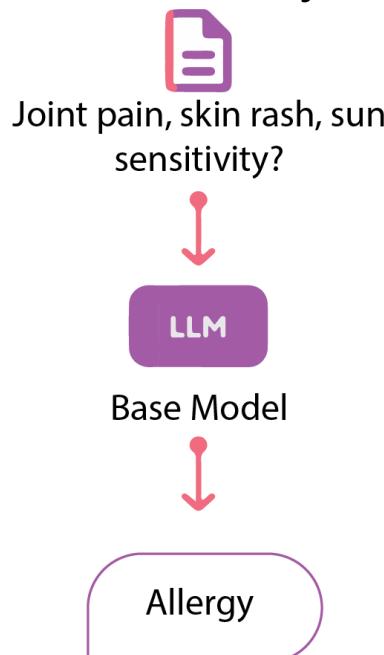


Lack of context and empathy



Why we need LLM Fine-tuning

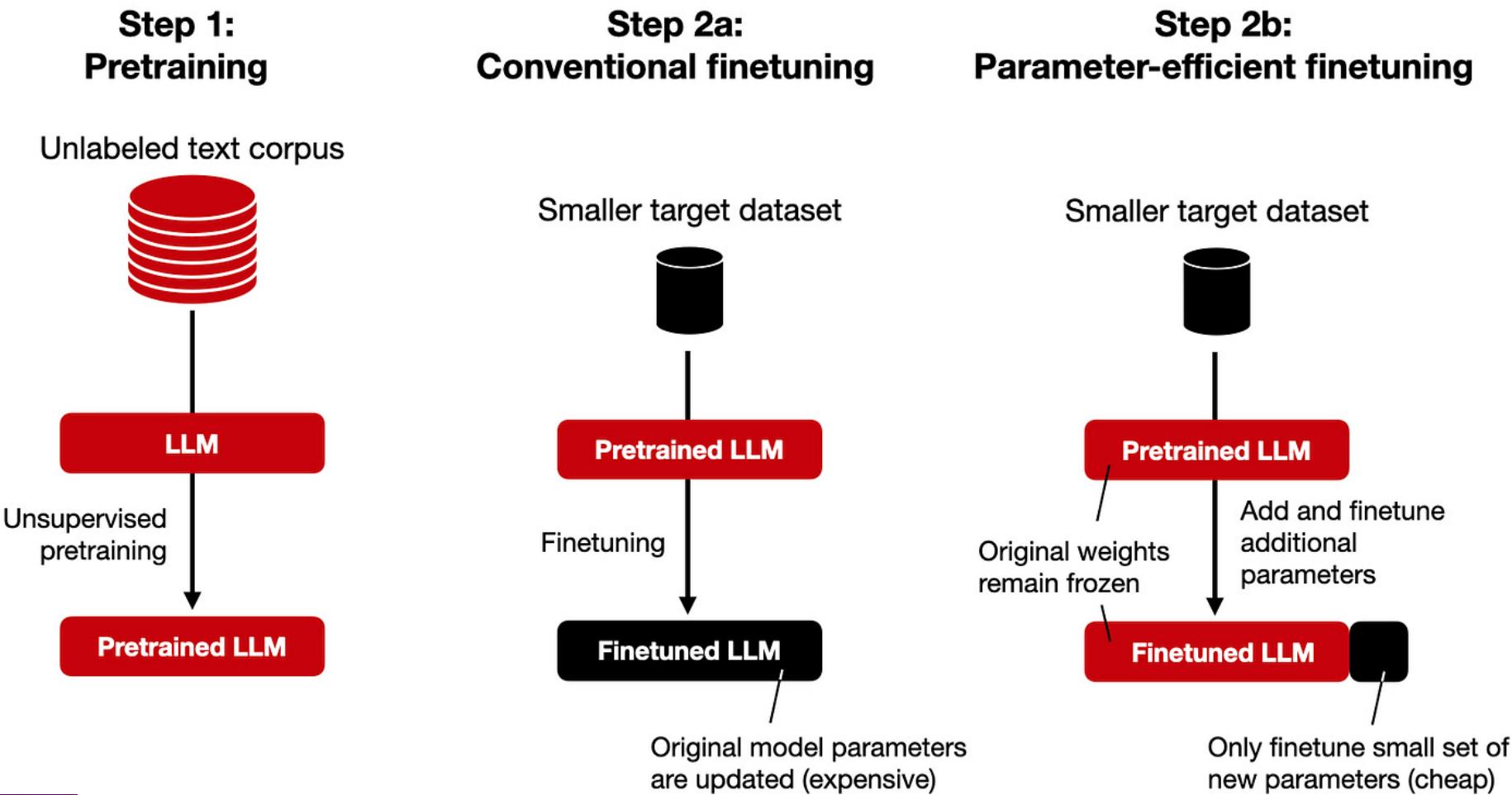
- Domain-Specific Adaptation: Tailoring a pre-trained LLM to specialize in a particular field, such as medicine.
- Focused Training Data: Uses specific datasets relevant to the target domain to improve accuracy and relevancy.
- Enhanced Performance: Improves the model's ability to generate precise, contextually appropriate responses in specialized areas.
- Reduces Irrelevant Output: Minimizes the risk of producing irrelevant or incorrect information by narrowing the model's focus.



Could be systemic lupus emphysematous (SLE), an autoimmune disease. It's vital to visit a dermatologist for a thorough examination and potential tests.

Fine-tuning Strategies

- After pretraining we can fine-tune LLM with smaller datasets via conventional or parameter-efficient methods.



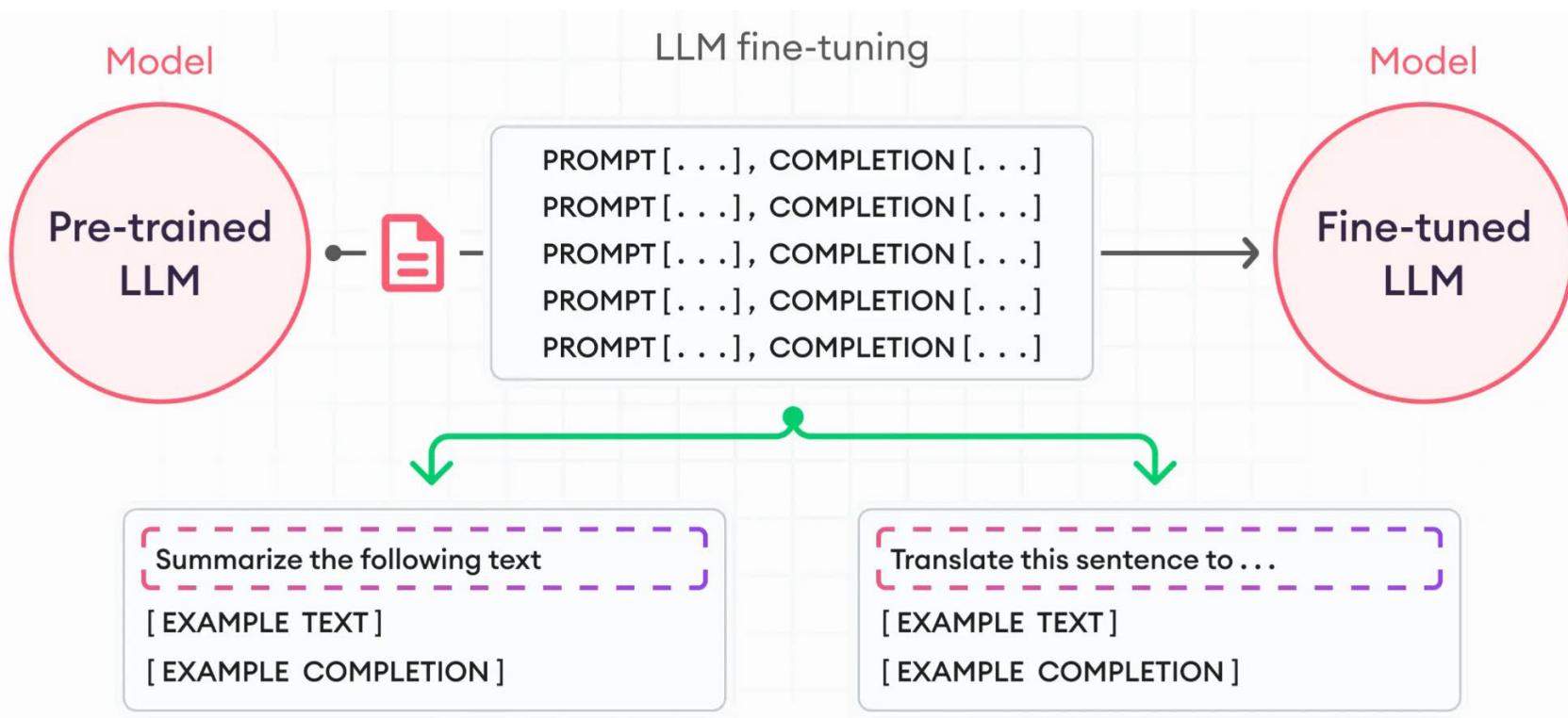
Cost for Pre-training and Fine-tuning

- Pre-training requires significant resources in terms of time and cost to be completed, whereas fine-tuning can be performed by an end-user with limited resources.

	Pre-training	Fine-tuning
Training time	Weeks	Hours
Compute	Thousands of GPUs	One or few GPUs
Dataset	Terabytes (e.g., C4 , Pile)	100-1000 MB
Budget	\$ Millions	\$ Hundreds

Instruction fine-tuning

- Instruction fine-tuning: Enhances model performance by training with examples that demonstrate desired responses.
- Task-specific datasets: Use prompt-completion pairs tailored to the task (e.g., summarization, translation) for targeted improvement.

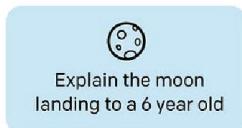


Supervised fine-tuning

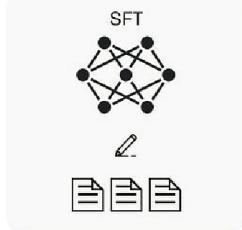
Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.

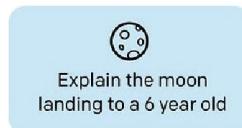


This data is used to fine-tune GPT-3 with supervised learning.

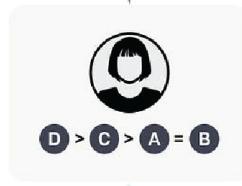
Step 2

Collect comparison data, and train a reward model.

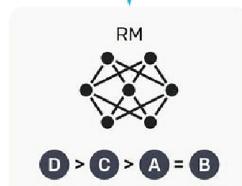
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



D > C > A = B

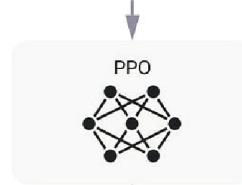
Step 3

Optimize a policy against the reward model using reinforcement learning.

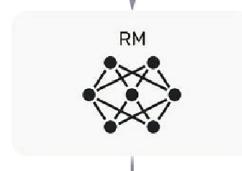
A new prompt is sampled from the dataset.



The policy generates an output.



Once upon a time...



r_k

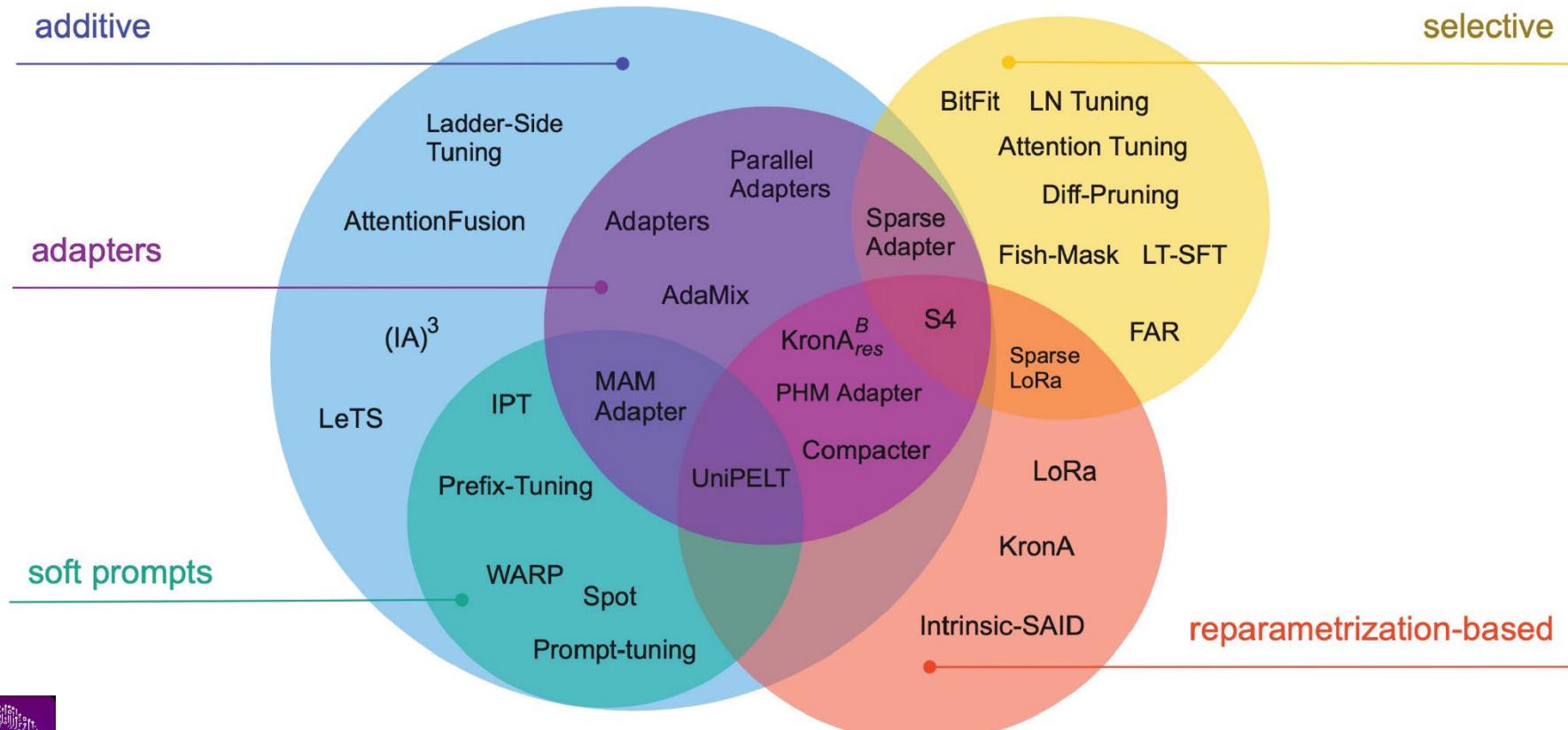
The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.



Parameter-efficient fine-tuning methods

- There are three main classes of fine-tuning methods:
 - Addition-based, Selection-based, and Reparametrization-based.
 - Within additive methods included groups like: Adapter methods and Soft prompts.



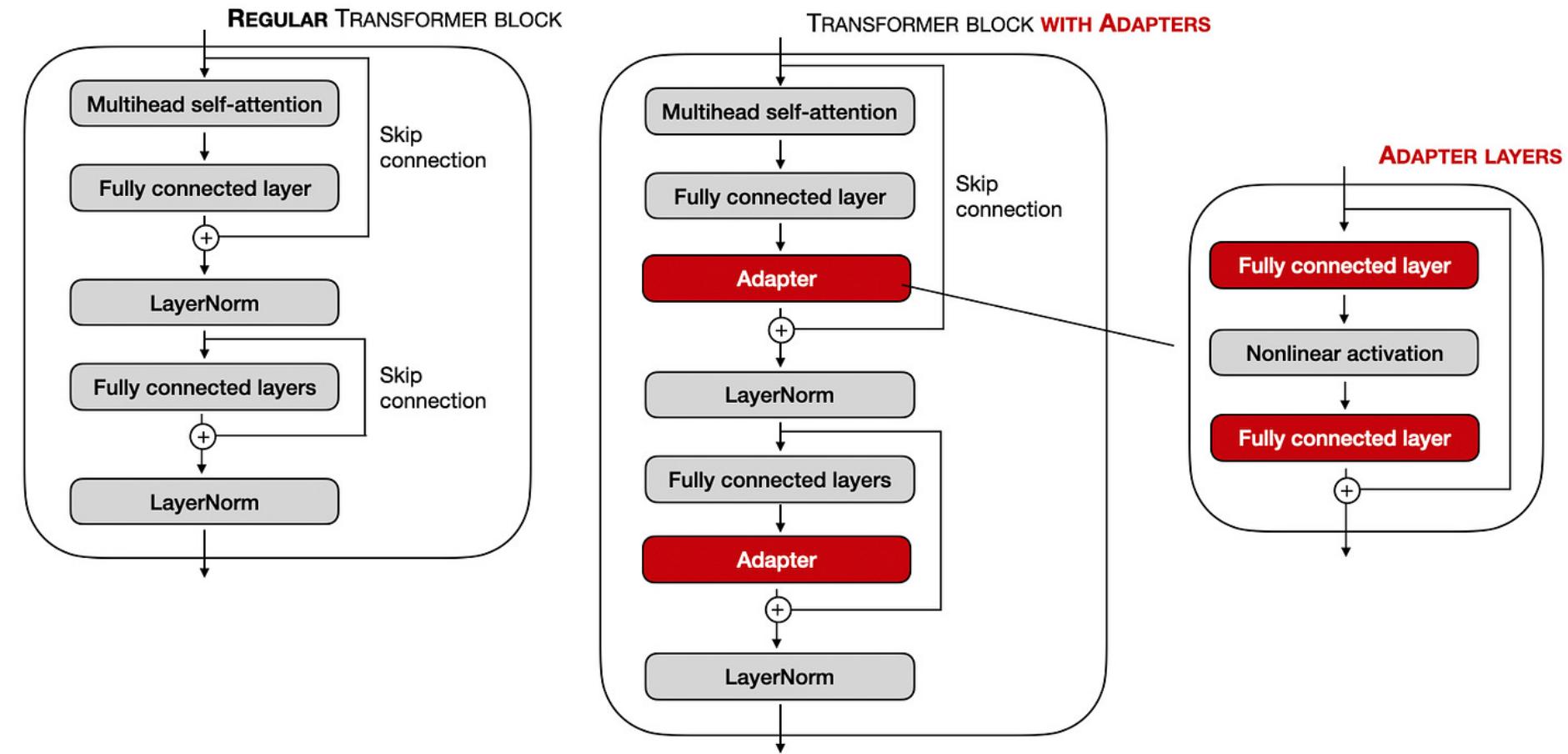
Fine-tuning taxonomy

- ❑ Full fine-tuning continues the initial training of the model using the existing checkpoint.
- ❑ PEFT (Parameter-Efficient Fine-Tuning) methods only fine-tune a small number of (extra) model parameters—significantly decreasing computational and storage costs—while yielding performance comparable to a fully fine-tuned model.
 - ❑ LoRA (Low Rank Adaptation) is a way to train large models efficiently by inserting (typically in the attention blocks) smaller trainable matrices to be learned during fine-tuning.
 - ❑ Prompt-based methods (p-tuning, prefix tuning, prompt tuning): Instead of manually creating hard (text) prompts, soft prompting methods are applied by adding learnable parameters to the input embeddings that can be optimized for a specific task while keeping the pre-trained model's parameters frozen.



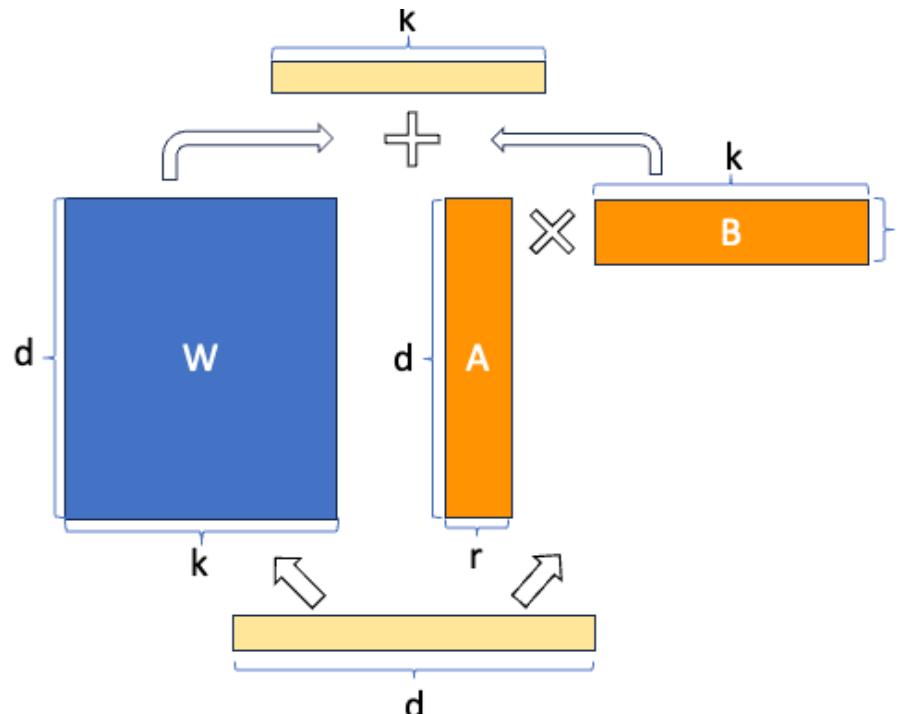
Adapter based fine-tuning method

- Comparison of the regular transformer blocks used in various LLMs and a transformer block modified via the adapter layers.



Low Rank Adaptation based fine-tuning

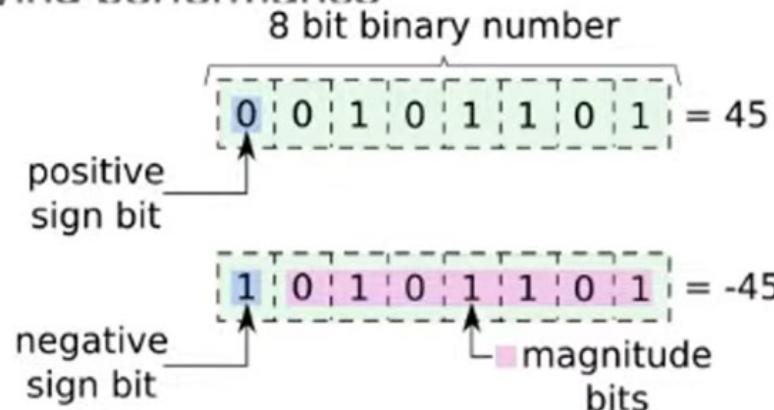
- LoRA fine-tunes two smaller matrices (A and B) that approximate the update to the original weight matrix (W_0) of the pre-trained LLM, rather than fine-tuning the entire matrix. The relationship is given by:
 - $W_0 + \Delta W = W_0 + BA$, where W_0 (dk), A(dr), B(r*k), and $r \ll d, k$.
- During training, W_0 is frozen while A and B are trainable. Initially, A is randomly initialized, and B is set to zero, making $\Delta W = BA$ start at zero. The matrices A and B form the LoRA adapter, allowing efficient fine-tuning with fewer parameters.



Low Rank Adaptation based fine-tuning

- Reduce computational costs (e.g., GPU memory needed) and preserve performance by **freezing model parameters** and using rank decomposition matrices to **cut down on the number of trainable parameters**.
- QLoRA: Efficient fine-tuning method for quantized LLMs that reduces computational costs by freezing model parameters and using low-rank adaptations.
 - QLoRA Optimization: Maps higher precision weights to lower bit spaces, preserving model performance while minimizing memory usage.

Map 16 bit integers into 4 bit space while preserving performance



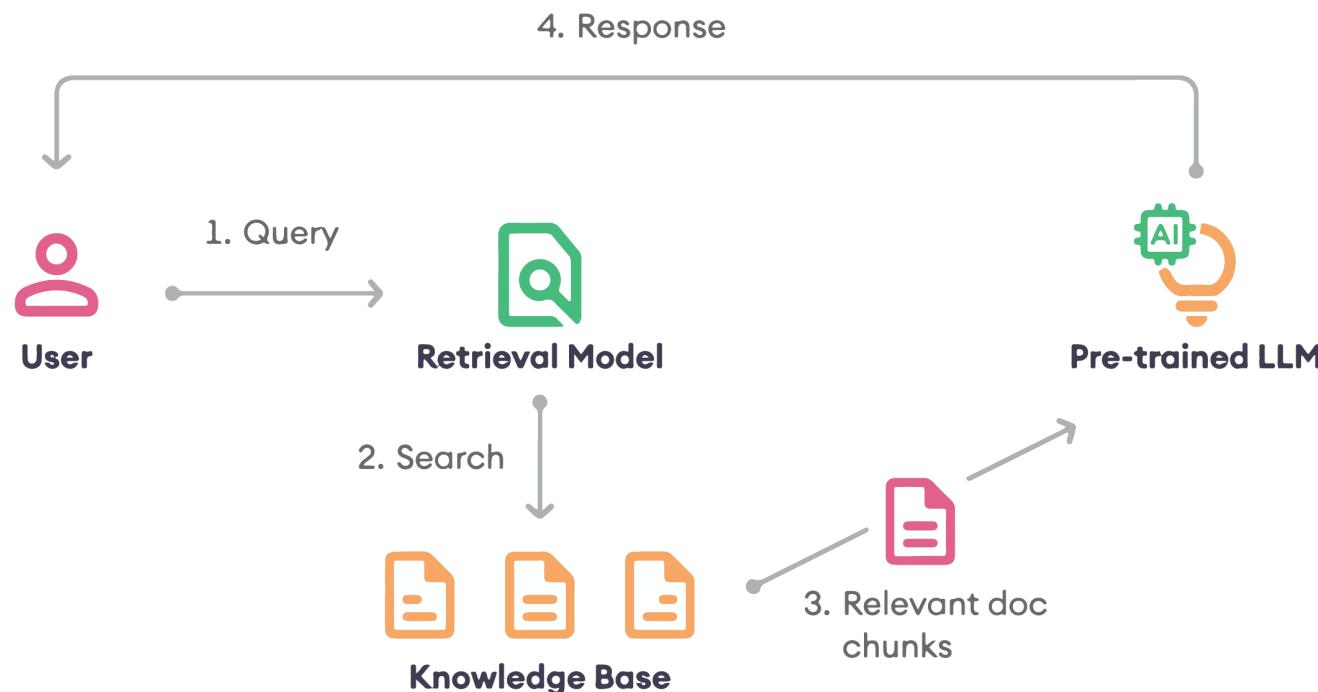
Methods to overcome LLM Challenges

- ❑ Inference performance optimization by reducing memory footprint and improving parallelizability.
 - ❑ Efficient attention using lower-level hardware-aware optimizations (e.g., Flash Attention).
 - ❑ Quantization by reducing the computational precision of weights and activations.
 - ❑ Mixture of Experts to decrease inference time by not using all experts at once.
- ❑ Hallucinations solutions
 - ❑ Retrieval-augmented generation (RAG)
- ❑ Misleading behavior solutions
 - ❑ Reinforcement Learning From Human Feedback (RLHF)



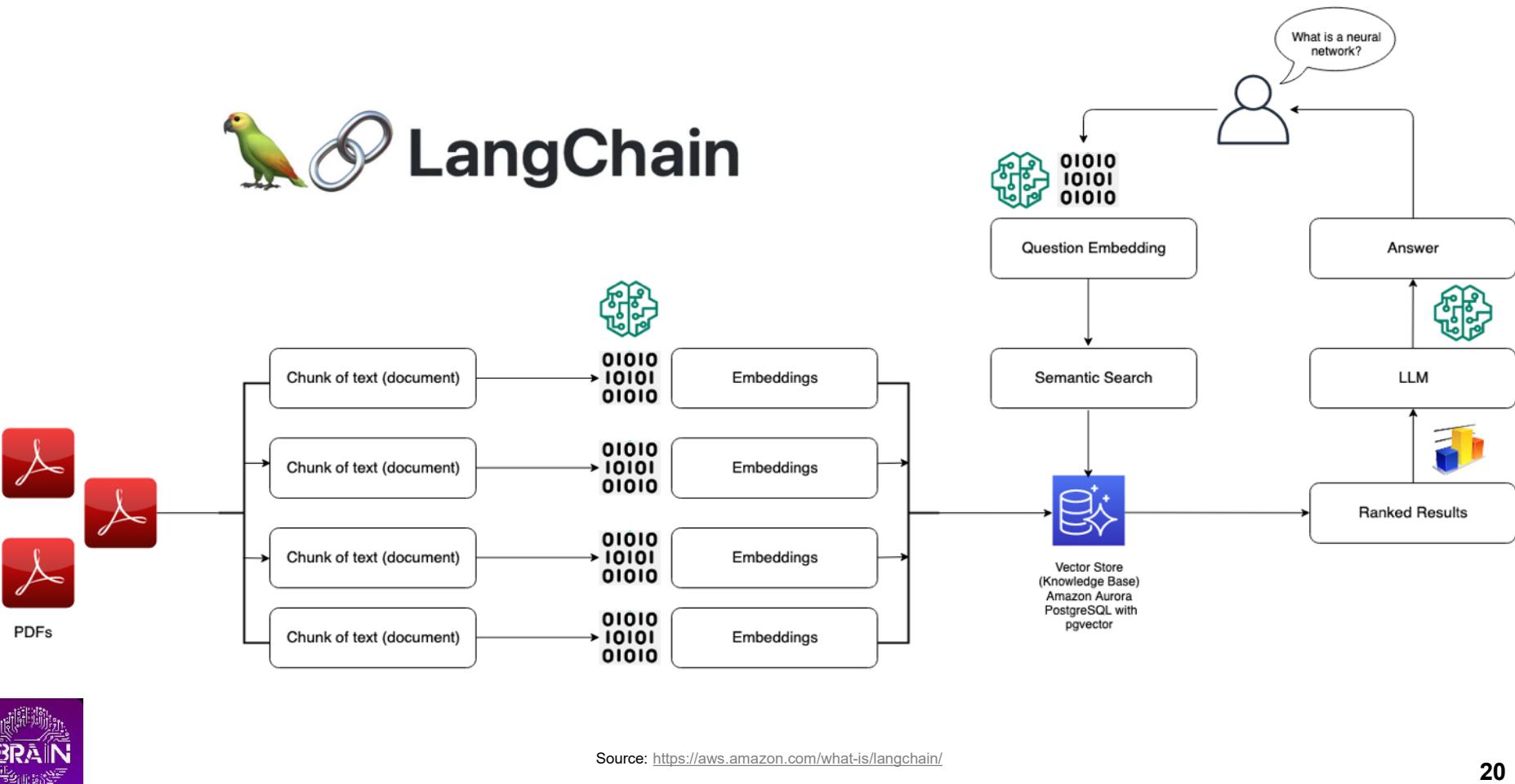
Retrieval Augmented Generation (RAG)

- RAG Overview: Combines natural language generation with information retrieval, grounding models in up-to-date knowledge.
- Dynamic Updates: Allows continuous data updates, unlike fine-tuning, which embeds static information.
- Complementary Techniques: Fine-tuning can enhance RAG systems by improving specific tasks, leading to better overall performance.



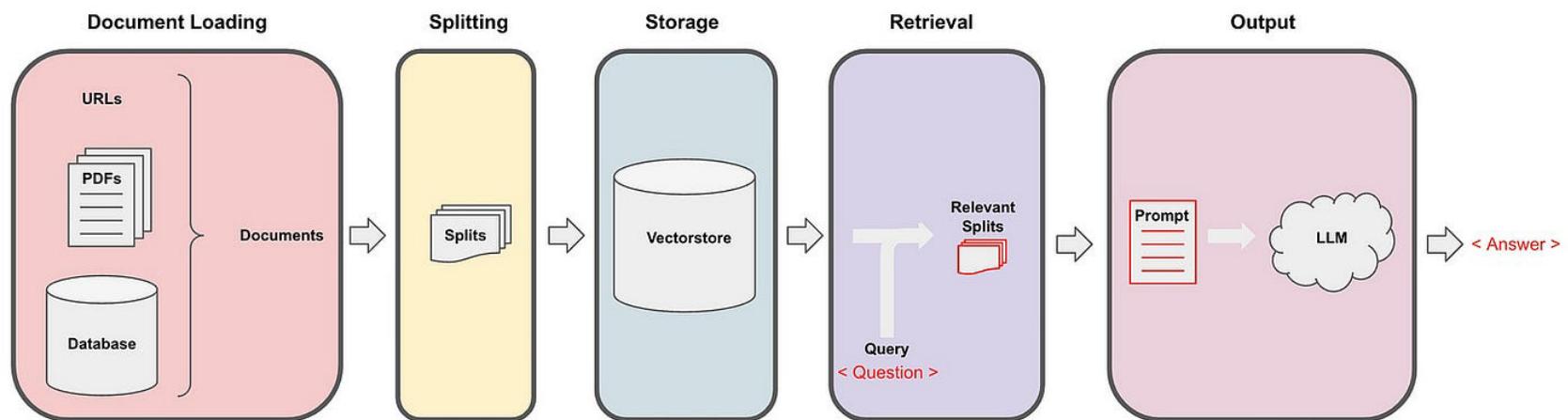
LangChain

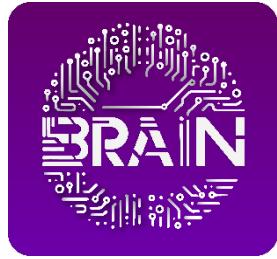
- ❑ Framework for building LLM Applications.
- ❑ Focus on retrieving data and integrating with LLMs.
- ❑ LangChain provides APIs with which developers can connect and query LLMs from their code.



LangChain Components

- ❑ LangChain enables developers to adapt language models to specific business needs by defining a sequence of steps (chains) to achieve desired outcomes.
- ❑ **Chains:** Chains in LangChain are sequences of automated actions that connect various AI components to handle tasks like querying data sources, generating content, and translating languages.
- ❑ **Links:** Links are the individual actions within a chain, allowing developers to break down complex tasks into smaller steps, such as formatting input, querying an LLM, or retrieving data.





MedAide Project

MedAide Overview



On-Premise Healthcare Chatbot

Designed specifically for hospitals and clinics, MedAide is an on-premise healthcare chatbot powered by consumer-end GPUs.



Human-friendly Conversational Style

MedAide's conversational style is tailored to be more human-friendly compared to other competitors, thanks to its training on healthcare forums.



Extensive Training on Healthcare Data

MedAide has been trained on millions of data points from online healthcare forums, enabling it to provide accurate answers to patients' questions about symptoms, prescriptions, and care plans.

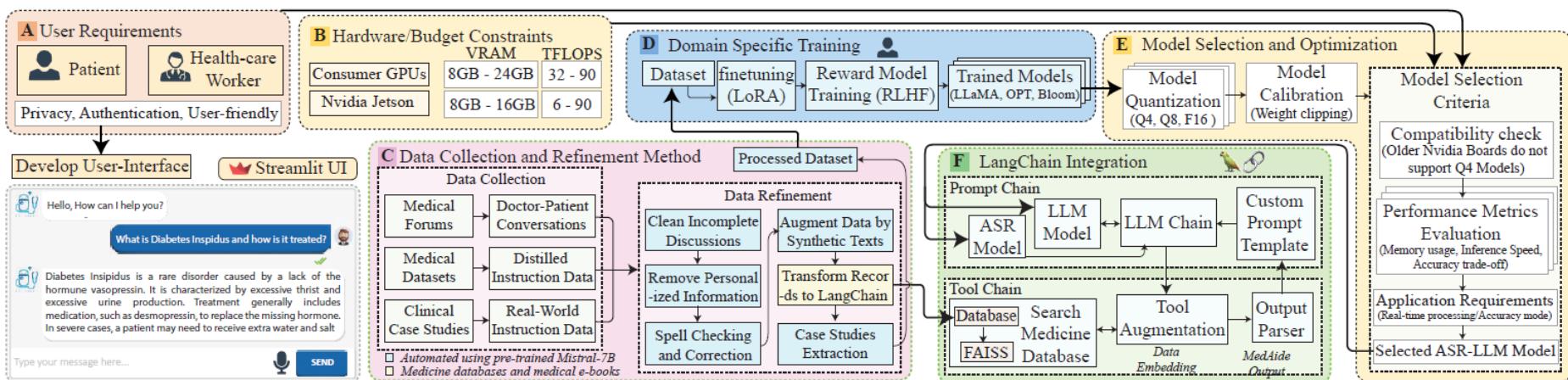


Optimal Performance and Resource Efficiency

With significantly fewer trainable parameters and optimized GPU memory usage, MedAide ensures outstanding performance while maximizing resource efficiency.

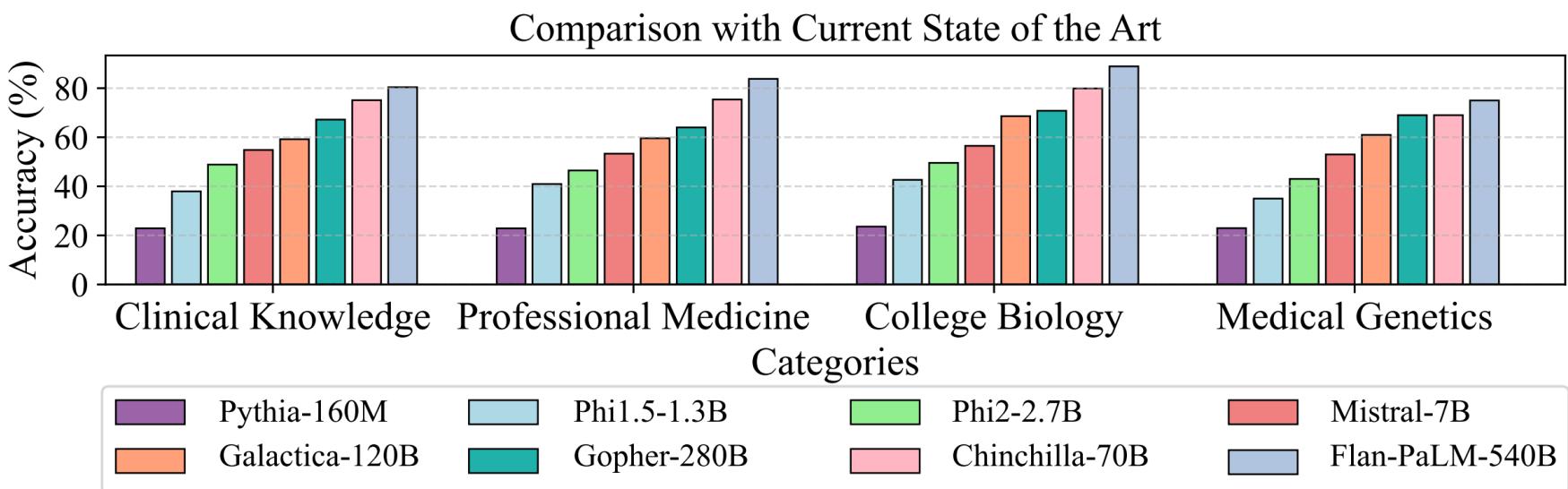
MedAide Project

- MedAide is an on-premise healthcare robot using LLMs integrated with LangChain backbone, ASR models, and model optimizations for efficient edge-based medical diagnostics.
- MedAide, deployed on Nvidia Jetson, achieves 78% accuracy in medical consultations and scores 56 in USMLE benchmark, offering resource-efficient healthcare assistance while addressing privacy concerns.



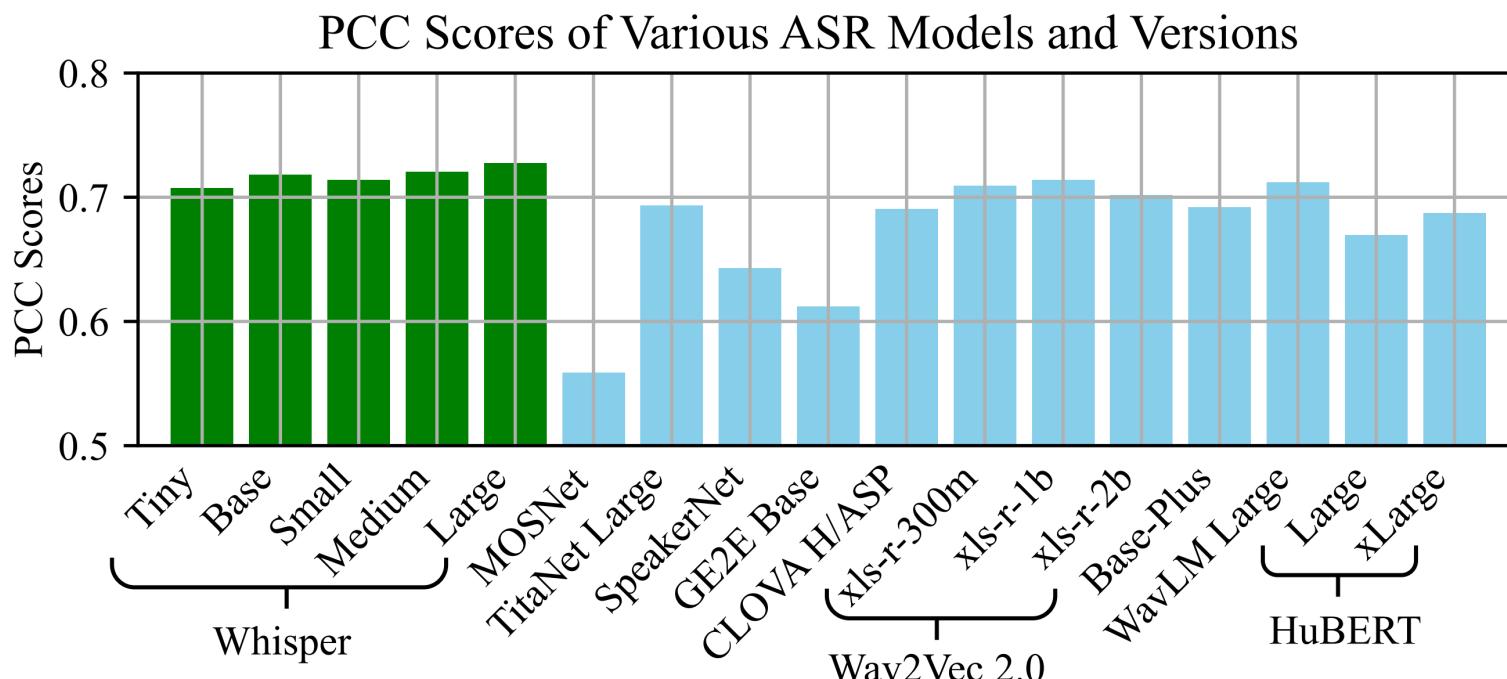
Comparison with State of the Art

- A comprehensive comparison between selected LLMs (Pythia, Phi1.5, Phi2, and Mistral) and state-of-the-art LLMs (Galactica, Gopher, Chinchilla, and Flan-PaLM) to evaluate the performance of these models on different domains, shedding light on their feasibility for medical assistance.



Comparison of different ASR Models

- ❑ Whisper models (Tiny to Large) consistently achieve higher PCC scores compared to other ASR models, indicating superior performance in this Automatic Speech Recognition (ASR).
- ❑ Diverse Performance: ASR models like Wav2Vec 2.0 and HuBERT show varied PCC scores, with some versions performing closer to Whisper, while others lag behind, highlighting differences in model architectures and training.



Training Dataset

- A comprehensive comparison between selected LLMs (Pythia, Phi1.5, Phi2, and Mistral) and state-of-the-art LLMs (Galactica, Gopher, Chinchilla, and Flan-PaLM) to evaluate the performance of these models on different domains, shedding light on their feasibility for medical assistance.



Kaggle / Research Datasets

Open source datasets on kaggle, research websites and HF such as GenMed and HealthCareMagic

Subreddits

We scrapped subreddits such as r/AskDocs and chose the top responses from certified doctors as a response

Medical eBooks

Widely used eBooks for USMLE and other top medical qualifications were used for RAG as well as finetuning the model

Healthcare News Websites

We also used healthcare news websites such as WebMD and HealthTap to scrape healthcare related datasets

Medical Forums

Forums such as Icliniq and eHealthForums were also used to gather accurate medical diagnosis

Certified MD Exams

Verified MD exams such as USMLE were also used to train as well as evaluate the model's accuracy on medical queries

Medical Datasets

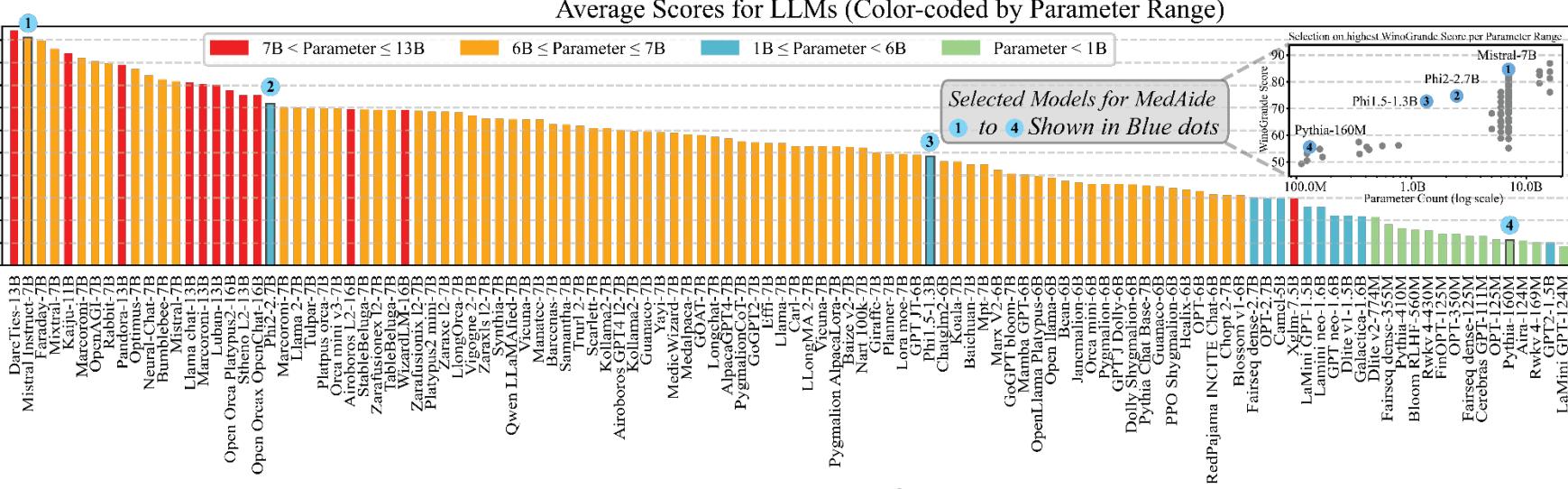
- The table lists various datasets used in medical question answering (QA), paraphrasing, and image-based QA, highlighting their domain, size, and data type, essential for training and evaluating AI models in healthcare.

Dataset	Domain	Size	Data Type
MEDQUAD	QA	11,230 questions	Clinical questions and answers
MEDMCQA	QA	1310 question–answer pairs	Clinical questions and answers
MEDQA-USMLE	QA	500+ question–answer pairs	Clinical vignettes and questions
MQP	QA	1000 question pairs	Clinical question pairs
CLINIQA-PARA	Paraphrasing	8130 paraphrase pairs	Clinical texts
VQA-RAD	Image QA	6000 question–image pairs	Radiological images and questions
PATHVQA	Pathology image QA	Over 14,000 image–question pairs	Pathology images and questions
PUBMEDQA	Abstract QA	Over 13,000 question–abstract pairs	Biomedical abstracts and questions
VQA-MED-2018, 2019, 2020	Visual QA	Over 8000 image–question–answer triples each year	Medical images, questions, and answers
RADVSDIAL	Conversational agent for radiology	Over 1500 dialogues	Text-based dialogues in radiology domain

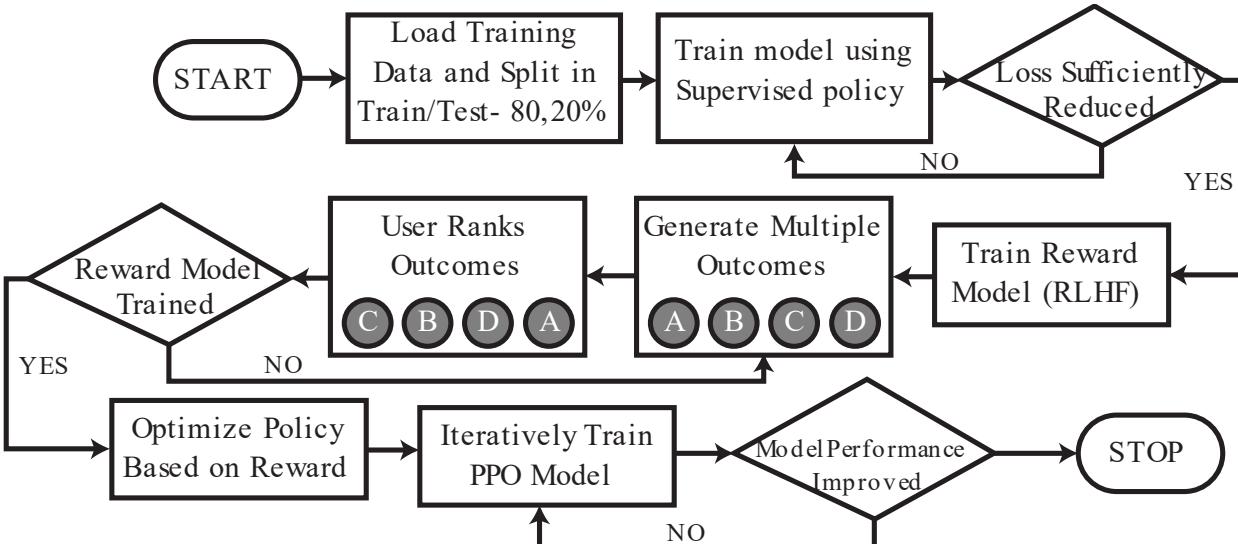
Model Selection and Training

- Various state-of-the-art LLM benchmarks on Open-LLM Leader-board including TruthfulQA, MMLU, ARC, HellaSwag, and WinoGrande were used to evaluate and select models for MedAide.

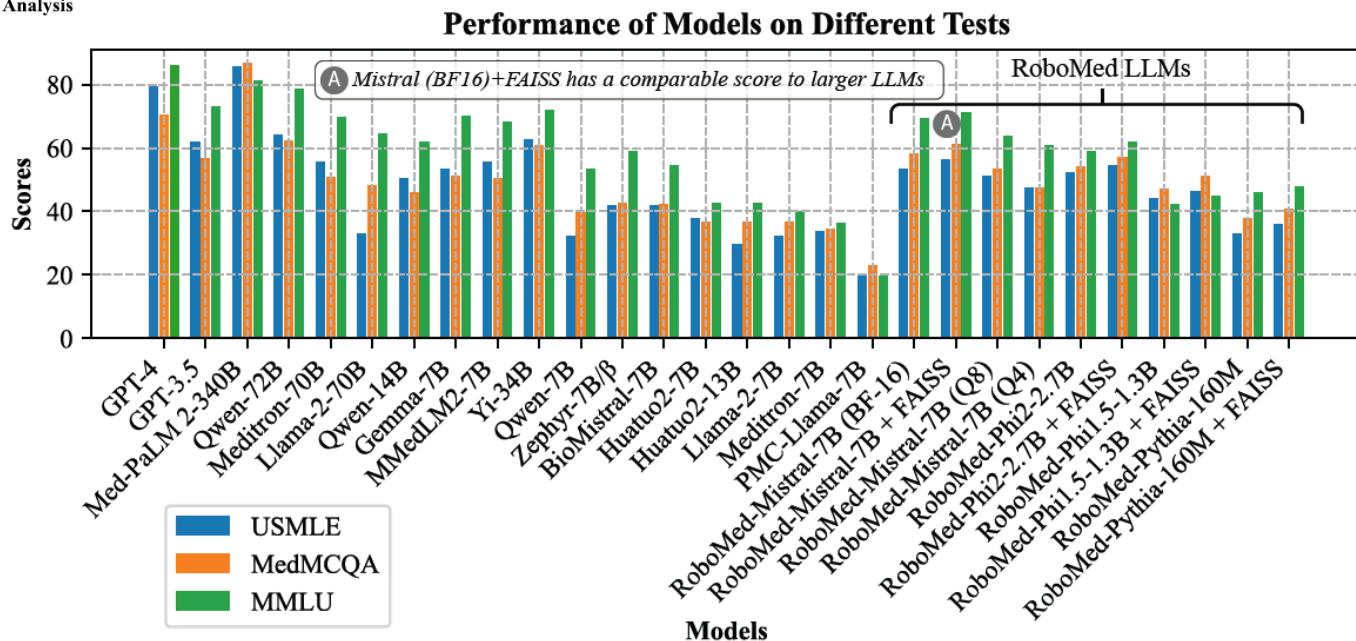
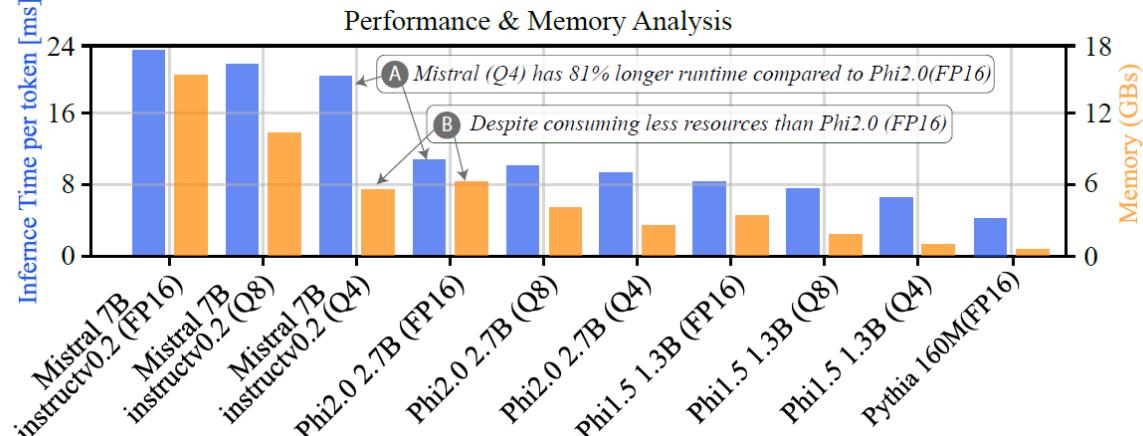
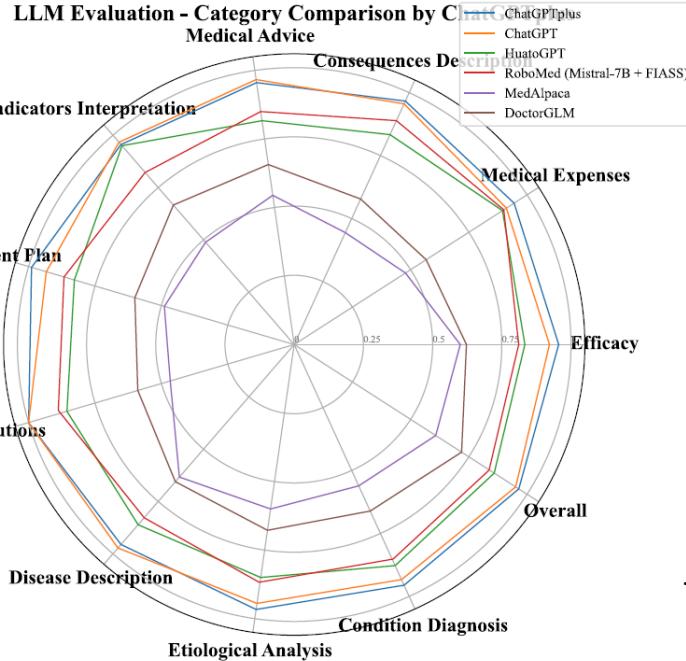
Average of {ARC, HellaSwag, MMLU, TruthfulQA} Benchmarks



Model Training Workflow



MedAide Project: Results



Comparison with similar Products

- MedAide Strengths: Offers comprehensive on-premise deployment, customizations, data security, and speech-to-text capabilities, making it a versatile and secure choice.
- Comparative Analysis: Drgupta.ai, Sensely, and FlorenceChat each lack in certain areas like on-premise deployment or speech-to-text, with FlorenceChat being a close alternative to MedAide in terms of features.

	MedAide	Drgupta.ai	Sensely	FlorenceChat
On-Premise	✓	✗	✗	✗
Customizations	✓	✗	✓	✓
Data Security	✓	✓	✓	✓
Speech to Text	✓	✗	✗	✓

Results Comparison with ChatGPT

Test 1:

- **Instruction:** Suggest treatment for pneumonia
- **ChatGPT Output:** The treatment for pneumonia typically involves antibiotics, rest, and supportive care such as staying hydrated and managing symptoms. It is important to consult a healthcare professional for a proper diagnosis.
- **MedAide Output:** Hi, it is advisable to take antibiotics like clarithromycin and levofloxacin tablets along with antipyretics like acetaminophen and anti-inflammatory drugs like ibuprofen. It is a good idea to take paracetamol (acetaminophen) tablets to relieve fever.

Results Comparison with ChatGPT

Test 2:

- **Instruction:** Is it okay to give Benadryl to a patient with asthma?
- **ChatGPT Output:** The use of Benadryl (diphenhydramine) in patients with asthma should be approached with caution. While Benadryl is an antihistamine that can help alleviate allergic symptoms such as itching, sneezing, and runny nose, it can also cause side effects that may affect individuals with asthma.
- **MedAide Output:** Benadryl treats allergies and it works for asthma as well. It also treats hay fever and many other allergic conditions. However, I would not recommend it as the sole treatment for asthma as it does not address the underlying causes and may not provide sufficient control of asthma symptoms. It is important to consult a healthcare professional for appropriate management of asthma.

Results Comparison with ChatGPT

Test 3:

- **Instruction:** I have a tight and painful chest with a dry cough, no fever and no headaches. Could it possibly be coronavirus?
- **ChatGPT Output:** While chest tightness, painful cough, and absence of fever/headaches can be symptoms associated with COVID-19, it is important not to jump to conclusions. These symptoms can also be caused by other respiratory conditions such as bronchitis or asthma. If you are concerned about COVID-19, it is advisable to get tested and consult a healthcare professional for an accurate diagnosis and appropriate guidance.
- **MedAide Output:** Hi dear, since you mentioned you do not have any fever, so you need not worry about coronavirus. But since you have painful chest with cough, you need to go for X-ray of the chest, which will help us to rule out bronchitis. You can also get a routine blood test to rule out any other issues like anemia or liver dysfunction. Regarding the medicines, you can take one course of antibiotics like tablet Cefixime.

Results Comparison with ChatGPT

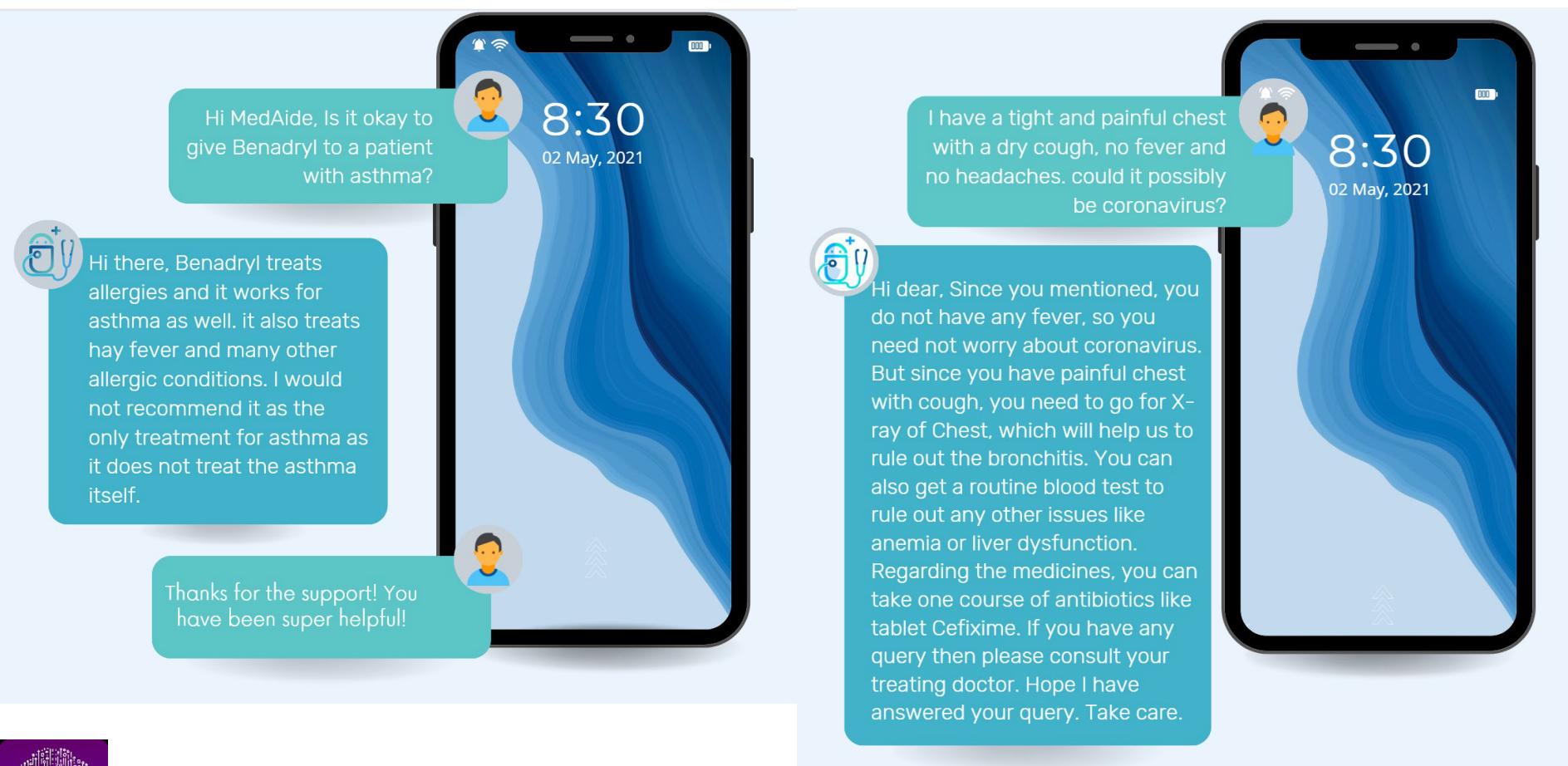
Test 4:

- **Instruction:** What are the fetal effects of maternal von Willebrand disease?
- **ChatGPT Output:** In general, von Willebrand disease (vWD) is a bleeding disorder that affects blood clotting. When it comes to the fetal effects of maternal von Willebrand disease, there can be some potential implications.
- **MedAide Output:** Weak hemostasis in the fetus can lead to intrauterine growth restriction, anemia, pre-term labor and birth, and fetal death.

Demo Video: MedAide

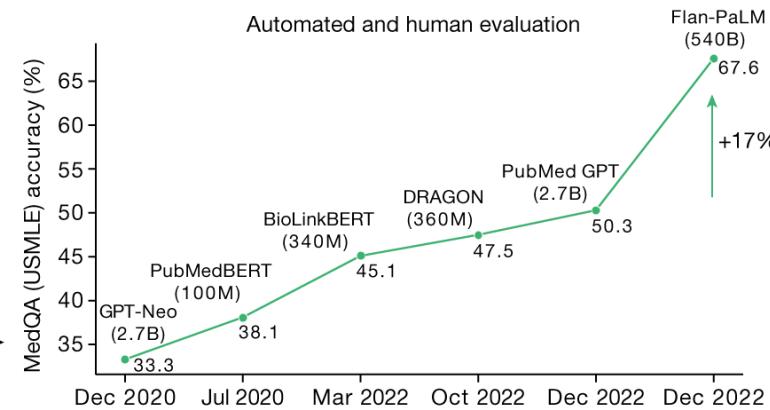
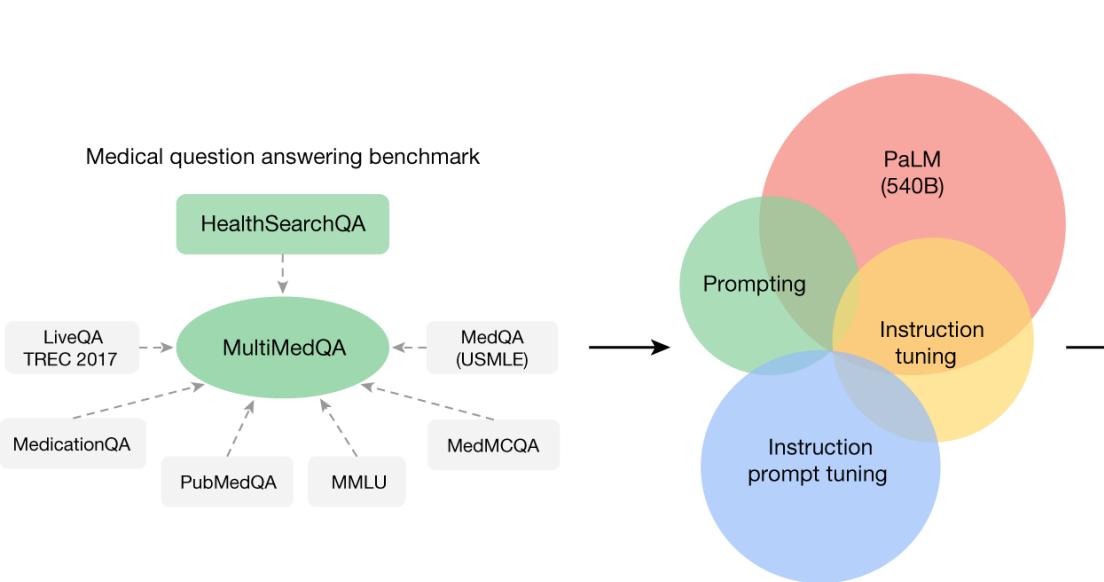
MindArm Deployment

- Our Goal is to deploy MindArm on a mobile phone device so that everyone has a personal medical companion on their phones.



Med-PaLM benchmark on MedQA Dataset

- MultiMedQA Benchmark: Integrates multiple datasets (e.g., MedQA, PubMedQA) to assess medical question-answering performance.
- Instruction Prompt Tuning: Combined with PaLM (540B), significantly boosts MedQA accuracy, reaching 67.6% by Dec 2022.
- Performance Improvement: Med-PaLM shows a 17% increase in accuracy, demonstrating potential in consumer medical question answering.



Q: How long does it take for newborn jaundice to go away?

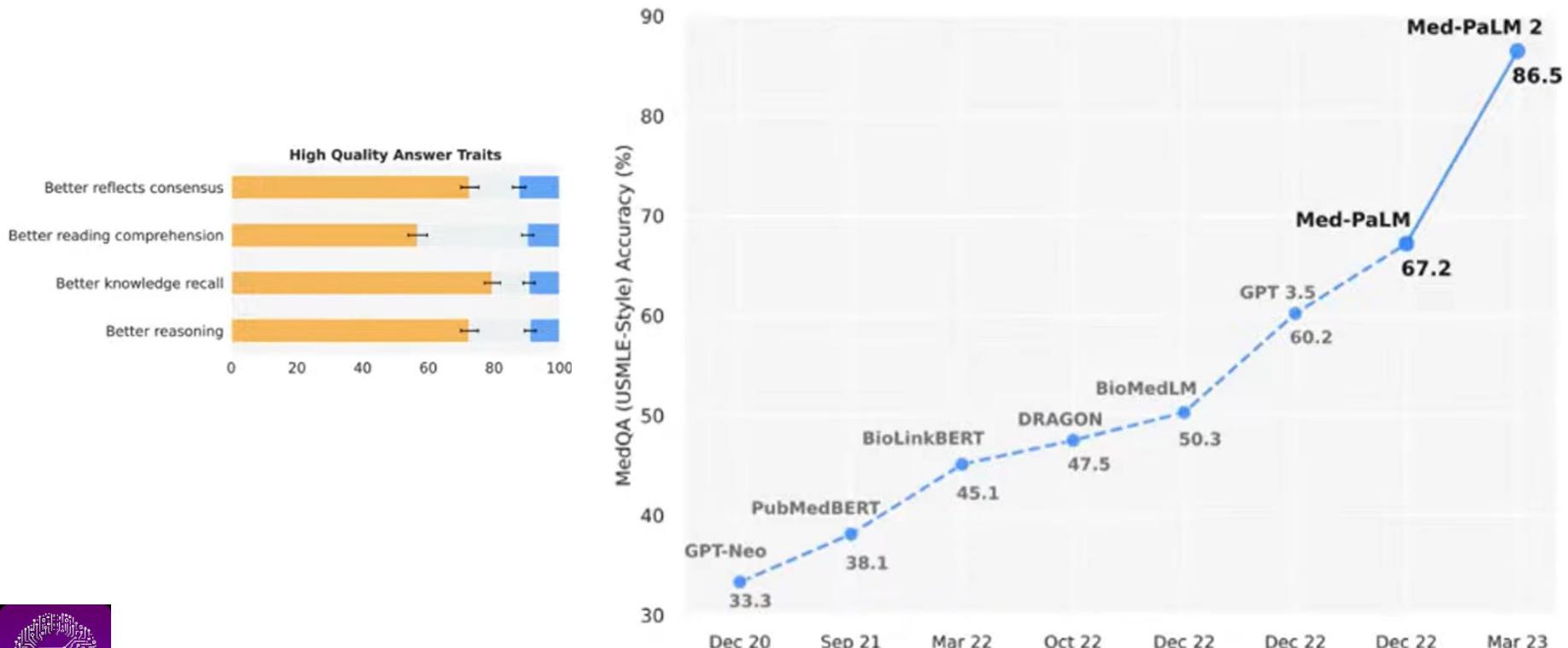
Med-PaLM: Newborn jaundice is when a newborn baby's skin and eyes look yellow. It is caused by a buildup of a substance called bilirubin in the baby's blood. Bilirubin is a byproduct of the normal process of red blood cells breaking down. Newborn jaundice is common and typically harmless, but it can be a sign of a more serious condition. The jaundice typically goes away on its own within a few weeks. In some cases, it may last longer or require treatment. The time it takes for the jaundice to go away can vary depending on the cause and the severity of the jaundice. If the jaundice is severe or lasts longer than a few weeks, the doctor may recommend testing or treatment to determine the cause and prevent complications.

Med-PaLM performs encouragingly on consumer medical question answering



Med-PaLM-2 on USMLE Benchmark

- MedPalm2 represents the cutting edge in medical language models, excelling in medical question answering tasks with a strong focus on accuracy and contextual understanding.
- Med-PaLM 2 significantly outperforms previous models, reaching 86.5% accuracy in MedQA (USMLE-Style) by March 2023, marking a substantial leap in medical question-answering capabilities..



Thank You!



abdul.basit@nyu.edu

#myNYUAD