RESEARCH ARTICLE

A Vision for SPARK Health: An AI Assistant Leveraging Biomedical FM/LLM for Smart Remote Senior Healthcare

Karthik SP Reddy

Temple City High School, 9501 Lemon Ave, Temple City, CA 91780; karthik.sp.reddy@gmail.com

ABSTRACT: The United Nations (UN) Decade of Healthy Ageing (2021-2030), led by the World Health Organization (WHO), is a global initiative to promote the well-being of seniors, families, and communities. The UN highlights that the global population of seniors is increasing substantially, and the WHO emphasizes the challenges this demographic shift poses for public health systems and social services, underscoring the need for inclusive digital solutions to support healthy aging. Remote healthcare teams, including family members and medical providers, often struggle to provide care for seniors due to geographic separation and the lack of integrated healthcare assistant systems, resulting in compromised quality of care. This work presents a conceptual proposal for Smart Personalized Assistant for Remote Knowledge on Health (SPARK Health), a personalized healthcare AI assistant designed to integrate with digital health platforms, leverage a biomedical foundation model (FM)/large language model (LLM), be managed by families, and facilitate collaboration with medical providers to address remote senior healthcare challenges. This paper reviews FMs/LLMs in healthcare and demonstrates how the conceptualized SPARK Health, leveraging a multimodal, multitask biomedical FM/LLM, could support health monitoring and healthcare decision-making and serve as a remote senior care solution aligned with the UN Decade of Healthy Ageing.

KEYWORDS: Foundation Model (FM), Large Language Model (LLM), Remote Senior Healthcare, AI Assistant, SPARK Health.

Introduction

UN's Global Initiative:

The United Nations (UN) World Social Report 2023: Leaving No One Behind in an Ageing World estimates that the global population aged 65 and older will reach 1.6 billion by 2050, more than double the 761 million recorded in 2021.[1] The World Health Organization (WHO, 2024) emphasizes the urgent need to ensure the health and well-being of seniors, the challenges this demographic shift poses for public health systems and social services, and the importance of inclusive digital solutions to support healthy aging.[2] The UN Decade of Healthy Ageing (2021-2030), led by the WHO as its Decade Secretariat, is a global initiative to promote the well-being of seniors, families, and communities.[3] The Smart Personalized Assistant for Remote Knowledge on Health (SPARK Health) is envisioned to align with the UN Decade of Healthy Ageing by addressing the needs of seniors, their families, medical providers, and the healthcare systems.

Challenges in Senior Care:

Historically, multi-generational households have provided care for seniors. Over the past decades, family and senior living arrangements have changed across both developed and developing nations.[1] Families are often separated by distance due to factors such as employment and/or education. Despite this separation, concern for loved ones, especially seniors who need more healthcare support, remains strong. Medical professionals also face challenges in managing the health of elderly patients. Particularly when the patients are not under direct supervision or are receiving outpatient care. Remote care teams (including family members, primary care physicians, geriatricians, and specialists managing specific health conditions) face the challenge of managing seniors' health without being physically present. Physical distance hinders timely interventions, causes a sense of ineffectiveness among care teams, and compromises seniors' health. Therefore, there is a requirement to improve remote healthcare and promote the well-being, comfort, and dignity of seniors in their later years. This research is motivated by the needs and experiences of seniors and families. The central research question that guides this study is whether leveraging an FM/LLM by a family-managed healthcare AI assistant can effectively bridge gaps caused by practical limitations in senior care.

Digitalization, AI, and Healthcare:

Digitalization, particularly the advancements of Artificial Intelligence (AI), is transforming the medical industry and healthcare administration.[4] It also helps reliably bridge the distance between individuals. Aggregating, analyzing, and interpreting data manually is impractical, time-consuming, and limits adequate healthcare. AI systems leveraging Foundation Models (FMs) and Large Language Models (LLMs), developed using advanced Machine Learning (ML) and Natural Language Processing (NLP), have redefined the way health data is collected, analyzed, and used. AI technologies such as Generative AI (Gen AI), FMs, and LLMs can handle various tasks and improve information management, decision-making, and medical processes.[5] The WHO's 2024 webinar, "Promoting Healthy Ageing in a Digital World," emphasizes that digital health tools, such as wearable devices and remote monitoring

DOI: 10.36838/IJHSR89.193

systems, can support healthy aging.[2] AI-powered Remote Patient Monitoring (RPM) systems help healthcare providers work more efficiently by supporting remote monitoring for timely interventions. They support patients by reducing the need for hospital visits and admissions, and assist healthcare systems by easing the burden on limited resources.[4] While various existing systems offer provide medical services, to the best of my knowledge, they lack one or more of the following: senior-focused personalization, a family-managed platform, expert age-related wellness guidance, conversational abilities, interoperability with digital health platforms, or the comprehensive integration of FM/LLM capabilities for tasks such as multimodal data analysis, concise routine and prompt-based summary report generation, risk prediction, disease diagnosis, or treatment recommendation.

SPARK Health: A Family-Managed Healthcare AI Assistant:

This paper presents a conceptual proposal for SPARK Health (Smart Personalized Assistant for Remote Knowledge on Health), a family-managed, personalized healthcare AI assistant powered by an FM/LLM, designed to provide expert-developed, individualized guidance to support remote monitoring and customized care for seniors. Building on the concept of standard RPM systems, SPARK aims to go further by enabling collaborative senior health monitoring with families and healthcare providers, functioning as a family-managed platform rather than a traditional facility-based approach used in RPMs. It aims to leverage the capabilities of FMs/LLMS, while RPM systems generally rely on conventional ML models and algorithms and rarely use FM/LLM capabilities.[4] SPARK Health is envisioned to address remote healthcare challenges by integrating with digital health platforms and leveraging an FM/LLM to deliver personalized, proactive, conversational, and collaborative real-time remote healthcare support to seniors, managed by families with involvement from healthcare providers.

Paper Overview:

This paper systematically reviews AI technologies in healthcare, primarily focusing on FMs and LLMs; outlines the training processes and taxonomy of FMs/LLMs; reviews key experimental findings in biomedical AI; identifies the type of FM/LLM to be leveraged by a healthcare assistant system; recognizes the model evaluation criteria; highlights the limitations of existing healthcare assistant systems; and presents a conceptual proposal for SPARK Health, a family-managed, personalized healthcare AI assistant, including its system architecture for integration with digital health platforms and leveraging an FM/LLM to support remote senior health monitoring and healthcare decision-making.

Literature Review

1. Artificial Intelligence (AI) Technologies:

The integration of AI in healthcare is growing due to its potential to improve access to medical knowledge, support providers in delivering care, and help individuals make informed health decisions.[6] In radiology, rising medical imaging has increased radiologists' workload, which can contribute to diagnostic errors, reader inconsistencies, and professional stress and fatigue. In such cases, AI can assist radiologists. Experts predict that while they do not expect AI to replace medical specialists, providers who use it will eventually replace those who do not.[7]

Machine Learning (ML):

Modern AI systems rely on machine learning, a branch of AI that enables models to learn tasks from examples instead of following explicit instructions. For instance, a single generic ML algorithm, such as logistic regression, can power diverse applications by learning directly from data.[8]

Three main paradigms of ML include:

- Supervised learning: The model learns from labeled input-output pairs of training data and usually needs a large amount of annotated data to make accurate predictions.
- Unsupervised learning: The model analyzes unlabeled data to find patterns, structures, or relationships and is commonly used for anomaly detection.
- Self-supervised learning: The model identifies features that capture the underlying structure and semantics of the data by predicting parts of the input from other parts of the same input. This approach uses large amounts of unlabeled data and is primarily used to pre-train large models.[8,9]

Deep Learning (DL):

Deep learning, a subfield of ML, enables models to learn complex patterns from large-scale data, including raw multimodal inputs, without requiring extensive manual feature engineering, unlike traditional ML algorithms. Deep Neural Networks (DNNs) with multiple layers have progressed significantly.[8,10] The same DNN architecture, such as Convolutional Neural Networks (CNN), can be used for many applications. These networks extract higher-level features when trained on raw inputs (e.g., pixels).[8] Other DNN types include Recurrent Neural Networks (RNNs) and Transformers.[10] With the rapid increase in medical images and associated data, DL models are increasingly used to support clinical decision-making.[11]

Natural Language Processing (NLP):

The main goal of NLP, a subfield of AI focused on language, is to enable machines to read, write, and converse like humans.[12] Along with related fields such as text-to-speech (TTS), NLP helps computers comprehend and produce human language.[8] In healthcare, the Named Entity Recognition (NER) technique extracts information about a patient's health, while text summarization summarizes treatment outcomes.[13] NLP applications improve the accuracy of clinical documentation, extract structured information such as vital signs, convert unstructured Electronic Health Record (EHR) data into structured formats, assist clinical decision-making, detect adverse drug reactions, de-identify records for privacy compliance, and support voice-to-text transcription in clinical workflows.[5,13]

Computer Vision (CV):

The primary way living organisms understand their environment is through vision, and CV, a subfield of AI, enables machines to interpret and understand visual information. Vision FMs, often built on DL architectures, are trained to analyze and interpret medical images such as X-rays, CT scans, and pathology slides. These models apply to medical image classification, disease diagnosis, and abnormality detection. Vision FMs, such as Segment Anything (SAM), have high potential for advancing medical image analysis.[5]

Generative AI (Gen AI):

Generative AI, a subfield of AI, comprises algorithms and models that learn patterns and structures in large training datasets to produce new data that resembles the original data. Modern Gen AI techniques often rely on DL neural network architectures. Gen AI models such as GPT generate human-quality text, Sora produces videos from textual prompts, and DALL-E creates images from textual instructions. Gen AI models, including generative FMs and LLMs, are used in the medical domain for applications such as providing personalized recommendations and generating summary reports.[10]

2. FMs and LLMs:

Healthcare systems around the world are facing significant challenges in improving overall wellness, providing better caregiver experience, and reducing the cost of care. The aging population is one of the main factors driving the need for transformed healthcare delivery models.[14] FMs and LLMs are now playing a central role, with vast capabilities to enhance the effectiveness of clinical, educational, and research work.[12] Table 1 presents various healthcare models, modalities, applications, and their base models.

Foundation Models (FMs):

Developing AI systems based on the general class of models called foundation models is an evolving approach.[5,8] FMs rely on DNNs and self-supervised learning, and leverage advances in ML, NLP, CV, and other technologies. They enable advanced functionalities and can standardize the model (e.g., GPT-3). They can aggregate information from data of various modalities and can be quickly adapted to downstream tasks through in-context learning or few-shot fine-tuning.[8,15] BiomedGPT and Med-PaLM M are notable examples of FMs. Across a wide range of tasks, a single FM can perform zero-shot predictions.[16] In healthcare, FMs help predict outcomes and track disease progress, provide automated diagnostic assistance, identify abnormalities in medical images, detect rare conditions, and enable personalized treatment planning, monitoring, and improvement of medication adherence.[5]

Large Language Models (LLMs):

LLMs are Gen AI models designed to comprehend and generate human language across diverse NLP tasks.[10,17] As FMs, they are pre-trained on massive datasets and then fine-tuned for specific tasks.[18] These large models can predict or generate linguistic units, such as words, phrases, or sentences, based on context, using mathematical methods to generalize language rules and knowledge for text generation and prediction.[5,12,13] Traditional NLP models have limited ability to generalize across datasets because their training typically focuses on specific tasks. In contrast, LLMs can understand context, interpret complex data, and produce human-like text across many NLP tasks. These models use DNNs with billions of parameters and are trained on vast amounts of unlabeled text through self-supervised learning.[13] GPT families and LLaMA are typical examples of LLMs. LLMs have shown strong potential in medical record summarization, patient data analysis, diagnostic assistance, personalized treatment recommendations, and support for medical decision-making.[5,13]

2.1. Training FMs/LLMs:

- Data Acquisition and Curation: Trainers collect large-scale data from clinical records, medical literature, healthcare guidelines, and domain-specific knowledge and curate it into datasets.[8,19]
- Pre-training: Models are pre-trained on datasets using self-supervised or unsupervised learning. They develop the ability to predict the next word in a sequence or fill in missing words or phrases, which enables them to grasp contextual information and syntactic structures.[5,8,15]
- Fine-tuning: Models are further trained on smaller, targeted datasets to adapt to specific domains or downstream tasks such as text classification.[5,8,17] Full fine-tuning or Parameter-Efficient Fine-Tuning (PEFT) methods, such as adapter or prompt tuning, are applied as needed.[16]
- Flexible learning approaches:
  - o Zero-shot: Model generalizes to unseen tasks or classes without prior examples.
  - o Few-shot: Model adapts to new tasks or classes with only a few examples.
  - o In-context: Model uses context or examples provided during inference to adjust to new tasks.[17]
- Prompt Engineering: Prompts are crafted and optimized to instruct the model effectively. In particular, Chain-of-Thought (CoT) prompting guides models through multi-step reasoning processes.[17] During usage, LLMs respond to prompts by producing text that is logical and relevant to the context.[5]
- Evaluation and Deployment: Models are evaluated and deployed for real-world applications.[8] They enable human-AI collaboration and the development of unified architecture AI systems that are capable of analyzing multimodal data for multiple tasks.[15]

The flowchart in Figure 1 illustrates FM training on natural and remote sensing images for vision tasks, following a process similar to LLM fine-tuning.

DOI: 10.36838/IJHSR89.193

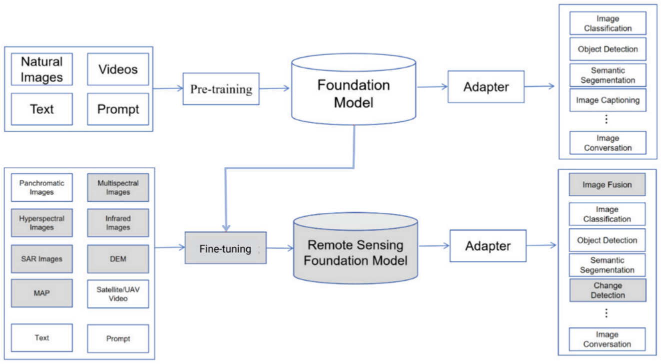

Figure 1: Flowchart of FM training on natural and remote sensing images for vision tasks, similar to LLM training. The model is first pre-trained on natural images and then fine-tuned using an adapter for various tasks. To handle remote sensing images, which differ significantly from natural images, such as in spectral context, the model is further fine-tuned and specialized using the adapter technique for additional remote sensing tasks.[16]

2.2. Taxonomy of FMs/LLMs:

This taxonomy provides a structured framework for comparing model capabilities and guiding model selection.

The FM/LLM classification categories include:

- Application: Specialized LLM and General-purpose LLM
- Modality: Multimodal LLM and Unimodal LLM
- Task: Multitask LLM and Single-task LLM
- Architecture: Encoder-only, Decoder-only, and Encoder-Decoder LLM
- Scope: Unified LLM and Modular LLM

Figure 2 depicts the above taxonomy.[12,15,17,20]

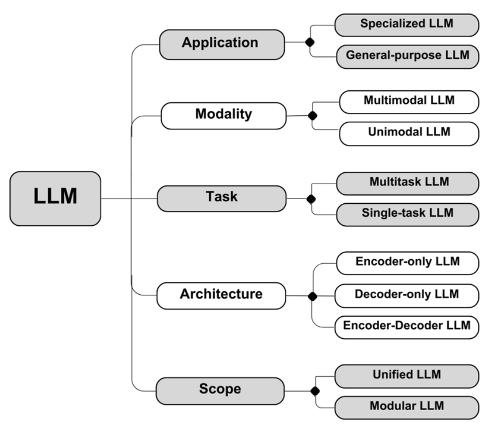

Figure 2: Taxonomy of LLMs depicting classification based on Application, Modality, Task, Architecture, and Scope, providing a structured framework to compare model capabilities and guide model selection.[12,15,17,20]

Application: Specialized LLM and General-purpose LLM:

LLMs can be categorized as specialized or general-purpose models based on their application.[19] General-purpose LLMs lack task-specific optimizations, whereas specialized LLMs are fine-tuned for a particular domain or task. Specialized models typically outperform traditional structured predictive models and can handle unstructured textual data such as EHRs.[17] For instance, PaLM-E achieved state-of-the-art (SOTA) performance across tasks; however, state-of-the-art (SOTA) performance across tasks. However, when evaluated on biomedical tasks in MultiMedBench, PaLM-E's performance is weak, and Med-PaLM M, an FM fine-tuned with domain-specific biomedical data, outperforms it.[15] The Autonomous ChatDoctor AI System utilizes ChatDoctor, a fine-tuned version of the general-purpose LLaMA model. This specialized model outperforms ChatGPT, a general-purpose model, due to significant improvements through fine-tuning for comprehending a patient's needs and providing informed advice.[5,17,20] These studies highlight the importance of fine-tuning to achieve strong performance on medical tasks and address the distribution shifts presented by the domain.[15]

Modality: Multimodal LLM and Unimodal LLM:

LLMs can be categorized as multimodal or unimodal models based on the modalities they can handle. Multimodality is particularly important in the medical field, which involves diverse data types, including text, medical images, audio, video, time series, sequences, and graph data.[5,15] Despite advances in biomedical AI, most models remain unimodal and handle only one data type.[15] For instance, image-based models are limited in interpreting textual or semantic information.[10] Patient data can be classified into three categories:

- Structured data: Quantifiable data, e.g., vital signs like blood pressure.
- Semi-structured data: Flow chart format data, e.g., lab test results.
- Unstructured data: Narrative-rich data, e.g., clinical notes.[13]

AI-enabled systems can record vital signs, physical activity, medication usage, and mental health parameters from wearable devices, smart home technologies, and mobile health applications for comprehensive monitoring.[4] These systems capture physiological data in real-time, including heart rate, sleep patterns, blood pressure, glucose levels, body temperature, respiratory rate, Electrocardiogram (ECG) signals, and activity metrics such as steps, distance, and calories burned.[4,21] For instance, PH-LLM can interpret time-series data from wearables (e.g., Fitbit and Pixel Watch) to provide insights and personalized recommendations, particularly in areas such as sleep and fitness.[22] Figure 3 presents comparative ECG flowcharts, illustrating the difference between unimodal and multimodal FMs/LLMs.[23]

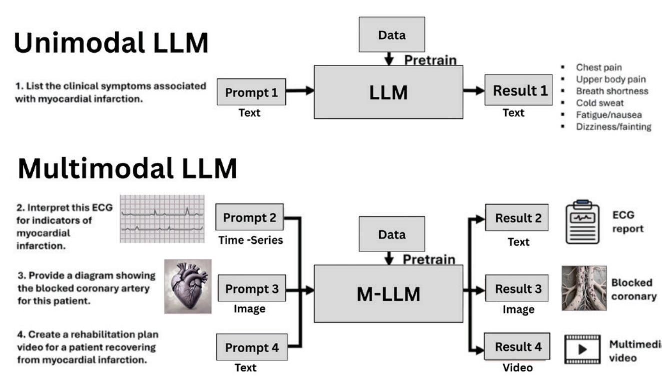

Figure 3: Flowcharts demonstrating Unimodal and Multimodal LLMs for ECG Analysis. The diagram depicts a Unimodal LLM generating output in the same modality as its input, while a Multimodal LLM processes multiple input modalities and generates outputs across various modalities.[23]

Task: Multitask LLM and Single-task LLM:

LLMs can be categorized as multitask or single-task models based on their task training. Although biomedical AI has advanced, most models are still designed as single-task systems that perform specific tasks. For instance, AI systems analyzing mammograms cannot explain their predictions verbally or interact in dialogue to incorporate a specialist's feedback, and outputs are limited to predefined classifications.[15] LLMs can achieve zero-shot generalization when pre-trained on unstructured text to perform multiple tasks without additional training. Pre-training uses a self-supervised objective rather than a task-specific one, enabling models to adapt to new functions without requiring additional data or training, and incorporating a multitask fine-tuning stage on prompted tasks after pre-training can improve their zero-shot abilities.[24] Med-PaLM M, BiomedGPT, and LLaVa-Med are examples of multitask biomedical FMs.[15]

Architecture: Encoder-only, Decoder-only, and Encoder-Decoder LLM:

LLMs are classified as encoder-only, decoder-only, or encoder-decoder based on their architecture.

- Encoder-only models (e.g., XLM-RoBERT) excel at contextual understanding tasks such as NER and text classification. However, they are less practical for text generation due to their limited generative capabilities and high computational needs for processing long inputs.
- Decoder-only models (e.g., GPT-4 and LLaMA 3.1) are strong in text-generation tasks. They support few-shot learning, producing context-aware responses from a few examples, and are suitable for content creation, conversational AI, and creative writing. However, they are limited in understanding context and need extensive computational resources to produce high-quality outputs.
- Encoder-Decoder LLMs (e.g., T5) combine the strengths of encoders and decoders, supporting comprehension and text generation tasks, including question answering and summarization. However, this dual architecture increases complexity and leads to training challenges and slower inference compared to other LLMs.[12,24]

Scope: Unified LLM and Modular LLM:

LLMs can be classified as unified or modular models based on their functional scope. Modular, task-specific architectures were developed earlier, with each module dedicated to a specific sub-task. In contrast, the modern approach involves using a single FM and adapting it using small amounts of task-specific annotated data, if necessary, to create an adapted model.[8] Med-PaLM M is an example of a unified FM that uses a single architecture with the shared model weights to perform a wide range of task types, including generating and summarizing radiology reports, medical question answering, and classification of medical images. PaLM-E is a unified generalist model that excels at language-only, vision-language, and embodied vision-language tasks. BiomedGPT and LLaVA-Med are additional examples of unified, generalist biomedical FMs.[15] By handling multiple modalities and functions across a wide range of domains, unified and generalist AI models provide a comprehensive view within a single architecture.[19]

Background

1. Review of Key Experimental Findings in Biomedical AI:

Visual FMs such as CLIP have achieved considerable success in CV and NLP. A comprehensive evaluation of MedCLIP, a medical multimodal model based on CLIP and fine-tuned on medical image classification tasks across four public datasets, shows that it outperforms the SOTA GLoRIA using only 10% pre-training data. Experiments also verify MedCLIP's transferability to various downstream tasks.[11,15]

Med-PaLM M, a unified biomedical multimodal generative FM, was evaluated on the MultiMedBench benchmark against specialist SOTA models and the generalist PaLM-E 84B (a baseline model without biomedical fine-tuning). Med-PaLM M used the same weights without task-specific adaptation across five task types: question answering, report summarization, visual question answering, report generation, and medical image classification. These tasks spanned modalities, including text, radiology (CT, MRI, X-ray), pathology, dermatology, mammography, and genomics. Med-PaLM M, fine-tuned with domain-specific data, outperformed PaLM-E and achieved or exceeded SOTA across tasks, datasets, and metrics.[15]

A scoping review of LLMs' potential in language understanding and processing in the EHR context, based on 329 papers collected from OpenAlex, was conducted by categorizing each paper into one of seven NLP tasks: NER, information extraction, text similarity, text summarization, text classification, dialogue system, and diagnosis and prediction. The results demonstrated that LLMs have significant potential to manage data, improve diagnostic suggestions, enhance patient engagement, and personalize medicine, provided issues such as data privacy and AI bias are being addressed.[13]

BiomedGPT, an open-source, lightweight, generalist vision-language FM, achieved SOTA results in 16 out of 25 benchmark experiments. In human evaluations of radiology tasks, such as visual question answering, report generation, and summarization, BiomedGPT demonstrated strong performance, with a 3.8% error rate in question answering, an 8.3% in writing complex radiology reports, and a summarization preference score nearly equivalent to human experts. These results indicate that training biomedical AI models on diverse data can improve practicality and clinical workflow efficiency.[28]

A study evaluated nine open-source generative LLMs on DRAGON, a Dutch clinical NLP benchmark, including 28 clinical information extraction tasks in Dutch. A publicly available framework, llm_extractinator, was developed on these LLMs and was used in a zero-shot setting to assess model performance. These LLMs proved highly effective for extracting medical information in Dutch, and when used with

DOI: 10.36838/IJHSR89.193

the framework, provide scalable, privacy-conscious solutions for low-resource clinical information extraction.[29]

2. Review of Existing Systems: Benefits and Limitations:

Telehealth monitoring was extensively utilized during the COVID-19 pandemic to maintain the safety of patients and providers, allowing patients to contact providers via audio or video using smart devices.[25] The Remote Patient Monitoring (RPM) systems are healthcare applications that allow medical providers to monitor patients' health outside clinical environments. RPM systems provide real-time health information for preventive care, including tracking physical activity, vital signs, and chronic conditions. These systems support continuous supervision, early detection, and timely interventions, which can help prevent complications, emergency room visits, hospitalizations, and readmissions, and reduce overall costs for seniors and healthcare systems. Hospital systems could reinvest these savings to improve access to medical services and healthcare resources.[4,25]

RPM systems often focus on patient health tracking and anomaly detection. Many are facility-centered rather than personalized for seniors or families. These systems rely on traditional ML models and algorithms and rarely leverage FM/LLM capabilities.[4] AI chatbots, such as Babylon and Ada, assist patients in recognizing symptoms and suggest potential further actions.[14] Emerald aids in remote monitoring of sleep, breathing, and behavior as a wireless, touchless sensor and ML platform.[14] Using motion and sound sensors, Google Nest monitors sleep and identifies disturbances like coughing.[14] CareTaker.ai, a smart health-monitoring and caretaker assistant, is designed for senior healthcare and supports automated monitoring, user interaction, and detection of health risks using an SVM model.[21] CaiTI leverages LLMs and sensor data for psychological therapy.[26] ChatGPT is a general-purpose conversational agent based on an LLM architecture, but it is not intended for specialized or personalized healthcare guidance.[27] Autonomous ChatDoctor AI System utilizes ChatDoctor, a fine-tuned version of the general-purpose LLaMA model to provide medical consulting and informed advice.[20] DrHouse is an LLM-powered virtual doctor system that supports multi-turn diagnosis by utilizing patient sensor data and expert medical knowledge.[26]

While various existing systems provide different medical services, to the best of my knowledge, they lack one or more of the following: senior-focused personalization, a family-managed platform, expert healthcare guidance, conversational abilities, interoperability with digital health platforms, or the comprehensive integration of FM/LLM capabilities for tasks such as multimodal data analysis, concise routine and prompt-based summary report generation, risk prediction, disease diagnosis, or treatment recommendation. There remains an opportunity for an integrated platform that unifies FM/LLM-powered capabilities and provides personalized, proactive, interactive, family-managed remote healthcare for seniors, accessible to seniors, families, and medical providers.

Methods

1. Model Type and AI Technique:

AI techniques have exhibited significant potential in performing diverse biomedical tasks, including interpreting radiology images, summarizing clinical information, and delivering accurate disease diagnoses.[28] Gen AI, FMs, and LLMs have enabled innovative healthcare solutions by improving efficiency in information management, decision-making, and medical process.[5] Gen AI has been identified as SPARK's core AI technique.

- The medical domain is highly specialized and critical, and FMs/LLMs explicitly trained for biomedical applications can provide more accurate and reliable support than general-purpose FMs/LLMs. Therefore, SPARK is designed to leverage a biomedical FM/LLM.
- Because healthcare data is fundamentally multimodal, SPARK would utilize a multimodal model rather than a unimodal one to process diverse data types effectively.
- Efficiently managing complex workflows and performing diverse healthcare tasks would require SPARK to employ a multitask model instead of a single-task one.
- A model built on an encoder-decoder architecture is expected to support comprehension, interpretation, and generation, rather than being limited to individual functions. SPARK would potentially use an FM/LLM with an encoder-decoder architecture.
- A unified architecture, in which a single model handles multiple data types and tasks, is preferred over modular models. SPARK would potentially employ an FM/LLM built on a unified architecture to process multimodal health data, including medical notes, images, audio and video files, and time-series data, and perform multiple healthcare tasks, such as analyzing multimodal data, generating concise routine and prompt-based summary reports, predicting risks, diagnosing diseases, and recommending treatments.

After examining the capabilities of FMs and LLMs to support the realization of SPARK, the type of FM/LLM that SPARK should leverage has been identified. A reliable, generative, multimodal, multitask biomedical FM/LLM, potentially built on a unified, encoder-decoder architecture, such as BiomedGPT, Med-PaLM M, and LLaVA-Med, adaptable to SPARK's focus on senior care, can be applied to achieve SPARK's objectives.

2. Dataset, Benchmarks, and Evaluation Metrics:

FMs/LLMs of the identified model type for SPARK must process multimodal health data, analyze and interpret inputs, generate accurate and reliable, multimodal, concise routine and prompt-based summary reports, predict risks, diagnose diseases, recommend treatments, and support remote monitoring of senior wellness. These models should have been pre-trained on diverse datasets, trained using Reinforcement Learning from Human Feedback (RLHF), and be adaptable to ensure strong performance on healthcare tasks, address distribution shifts presented by the domain, and maintain reliability in terms of quality, accuracy, and ethical standards.[12]

In recent years, there has been a surge in the development of FMs and LLMs for tasks such as NLP, CV, and healthcare due to the availability of extensive datasets and advances in DL.[5] Standardized metrics are employed to evaluate models and indicate the degree of similarity between generated and reference answers, typically ranging from 0.0 to 1.0. For instance, BLEU evaluates word and phrase overlaps between a reference and a model's output.[17] High-quality benchmarks have significantly contributed to the advancement of AI. MultiMedBench is a multitask, multimodal biomedical benchmark comprising 14 individual tasks across five task types and 12 de-identified open-source datasets.[15] HealthBench is an open-source benchmark developed in collaboration with 262 physicians across 26 specialties, with a collective practice experience spanning 60 countries, and containing 5,000 realistic health conversations between models and users.[6] In the medical domain, evaluation metrics, datasets, and benchmarks play a crucial role in evaluating the performance, reliability, multimodal, and Gen AI capabilities of the models.[6,15,23] FMs/LLMs would be evaluated using benchmarks such as HealthBench, MultiMedBench, DRAGON, and MultiMedQA; datasets such as MedQA, SLAKE, and PathVQA; and standardized metrics like F1-RadGraph and Accuracy.[17,28,30] Table 2 presents various datasets and metrics for evaluating healthcare models.

Depending on the evaluation results, the best-performing model for SPARK would be selected. If needed, the model may be fine-tuned and/or adapted for specific tasks and further improved through prompt engineering to enhance performance and enable efficient remote senior wellness monitoring.[16]

3. Tables:

Table 1 lists the modalities, applications, and base models of healthcare FMs and LLMs.[17] Table 2 presents various task types, domains/modalities, datasets, and metrics for healthcare FMs and LLMs.[15,28]

Table 1: Modalities, Applications, and Base Models of Healthcare FMs and LLMs.[17]

Model	Modality	Application	Base model
PathAsst	Pathology	Pathological diagnosis	PLIP, Vicuna-13B
BiomedGPT	Radiology, pathology	VQA, image captioning	OFA
PMC-VQA	Radiology, pathology, microscopy, etc.	VQA	PMC-CLIP, PMC-LLaMA
LLaVA-Med	Radiology, pathology	VQA	LLaVA, CLIP
XrayGPT	X-ray	VQA	MedCLIP, Vicuna
CephGPT-4	Dental imaging	Orthodontic measurement and diagnostic	MiniGPT-4, VisualGLM
Med-PaLM M	Radiology, pathology, mammography, genomics, dermatology, etc.	QA, VQA, report summarization and generation, genomic variant calling, etc.	PaLM-E
Med-Flamingo	Radiology, pathology, etc.	VQA	OpenFlamingo
RadFM	Radiology	VQA, disease diagnosis, report generation	ViT, PMC-LLaMA
BioMedGPT	Molecule, protein	QA (medical, molecule, protein)	LLaMA2-7B-Chat
HeLM	Individual-specific (ex. lab values)	Disease risk estimation	PaLM-E

Table 2: Task Types, Domains/Modalities, Datasets, and Metrics for Healthcare FMs and LLMs.[15,28]

Task Type	Domain/Modality	Dataset	Metric
Report Summarization	Radiology	MIMIC-III	ROUGE-L
			BLEU
			F1-RadGraph
Text understanding	Clinical notes	MedNLI	Accuracy
Clinical-Trial Matching	Clinical trials and patient's medical records	TREC 2022	
Text summarization	Doctor-patient dialogues	MeQSum	ROUGE-L
		HealthcareMagic	
	Radiology report	MIMIC-CXR	ROUGE-L
			F1-RadGraph
		MIMIC-III	ROUGE-L
			F1-RadGraph
Visual Question Answering (VQA)	Radiology	VQA-RAD	BLEU-1
			F1
		Slake-VQA	BLEU-1
			F1
	Pathology	Path-VQA	BLEU-1
			F1
Report Generation	Chest X-ray	MIMIC-CXR	Micro-F1-14
			Macro-F1-14
			Micro-F1-5
			Macro-F1-5
			F1-RadGraph
			BLEU-1
			BLEU-4
			ROUGE-L
			CIDEr-D
Image Classification	Chest X-ray	MIMIC-CXR (5 conditions)	Macro-AUC
			Macro-F1
	Dermatology	PAD-UFES-20	Macro-AUC
			Macro-F1
	Mammography	VinDr-MAMO	Macro-AUC
			Macro-F1
		CBIS-DDSM (mass)	Macro-AUC
			Macro-F1
		CBIS-DDSM (calcification)	Macro-AUC
			Macro-F1
	Genomics	Variant Calling	Indel-F1
			SNP-F1
	Colon pathology	NCT-CRC-HE-100K	Accuracy
	Dermatoscopy	HAM10000	
	Retinal OCT	Zhang Lab Data	
	Chest X-Ray		
		MC-CXR	
		SZ-CXR	
	Breast Ultrasound	Breast Ultrasound	
	Blood Cell Microscope	Blood Cell Microscope	
	Coronal abdominal CT	LiTS	
	Mass	CBIS-DDSM	F1-Macro
	Calcification		
Image Captioning	Chest X-Ray	IU-X-RAY	CIDEr
	Digital camera	PEIR-GROSS	
	Radiology	MIMIC-CXR	
Question Answering	Text	MedQA	Accuracy
		MedMCQA	
		PubMedQA	

Discussion

1. Overview and Applications of SPARK Health:

SPARK Health (Smart Personalized Assistant for Remote Knowledge on Health) is a conceptual, family-managed, personalized healthcare AI assistant designed to help families with health monitoring and medical providers in healthcare decision-making, to support seniors' healthcare remotely.

SPARK aims to empower seniors to adopt healthier lifestyles, follow treatment plans, and manage their health by tracking their status, setting personal goals, and receiving reminders for medications and activities. It enables monitoring and guidance for seniors in their environments, without physical contact. Since their health information would be maintained digitally, the inconvenience of managing paperwork can be avoided, and seniors and their providers could easily access their comprehensive health information from their devices during medical consultations. SPARK would provide personalized health guidance through individualized health tips, customized tasks, and condition-specific questionnaires tailored to

DOI: 10.36838/IJHSR89.193

each senior's unique needs, such as nutrition, exercise, biometric ranges, and medication, which would be co-developed by healthcare experts and the seniors' families. It would also provide general age-related tips for seniors, including measures to prevent diabetes, memory loss, high blood pressure, dementia, mental health issues, arthritis, and other conditions.

SPARK would enable real-time monitoring of seniors' health by care teams and empower them to stay updated despite physical distance. To achieve this, it would communicate with external digital health platforms to aggregate real-time wellness data of various modalities, and monitor parameters such as vital signs, activity levels, and more. It would also accept multiple types of input through the user interface to increase diagnostic accuracy. SPARK would automatically send alerts to care teams or emergency services when needed. It would generate concise routine, and prompt-based summary reports and allow Users to choose between textual and conversational input and output. SPARK would support comprehensive health history tracking by providing access to seniors' data and health history reports when required. To ensure continuous care, it would offer centralized, secure storage and management of data, enable access from any location or device at any time, and support simultaneous access by multiple Users.

SPARK would allow only healthcare providers to request risk predictions, disease diagnosis, and treatment recommendations, as such information may cause stress to seniors and their families, and any inaccuracies could have serious consequences.[20] This measure would ensure that providers verify all medical information before it is communicated to seniors or their families. SPARK would help providers work more efficiently by allowing them to focus on senior care and timely interventions, instead of spending time aggregating health information from various records, laboratories, and other sources. It would support providers in making decisions, minimizing professional stress and fatigue, and reducing the risk of diagnostic errors. It would ease the burden on the limited resources of the healthcare system and save costs for seniors, families, and healthcare systems.

SPARK aims to support individualized health tracking and reporting, keep all stakeholders informed through a shared, integrated platform, and help achieve significant outcomes in senior care.

2. System Architecture and Implementation:

The system architecture of SPARK is shown in Figure 4. SPARK would comprise a cloud-based user interface (UI), backend application, and database, and would interact with Users, external health data sources, and a generative, multimodal, multitask biomedical FM/LLM.[4,18,27]

Simulated Workflow:

Users (Stakeholders) → Data Acquisition (UI and Digital Health Platforms) → Data Preprocessing and Storage (Backend and Database) → Data Processing and Output Generation (FM/LLM) → Output Presentation and Notifications (UI display/voice-based, Alerts, Reminders) → Evaluation and Refinement → Deployment and Maintenance

Users (Stakeholders):

Primary Stakeholders would include seniors, their family members, primary care physicians, geriatricians, nutritionists, and specialists (e.g., cardiologists and neurologists), all of whom can access SPARK with role-based privileges granted with the senior's consent. Users would interact with SPARK through the UI for health monitoring and healthcare decision-making. Each user would have a login profile, and either the senior or a designated family member would be the chief user and manage permissions for other roles. Seniors may have limited access depending on their health and/or their family's judgment. SPARK supports multiple user profiles e.g., for households with multiple seniors.

Data Acquisition:

Users provide prompts, questionnaire responses, health records, and other inputs through the UI in text or voice formats, and in various multimodal health data types. SPARK would also aggregate data from multiple external health sources through application programming interfaces (APIs).

The data would be in diverse types, including text (e.g., medical notes), medical images (e.g., X-rays), audio files (e.g., heart sound recordings), video files (e.g., physical therapy exercises), time-series data (e.g., heart rate), sequence data (e.g., genomic sequences), and graph data (e.g., medical knowledge graphs).[5,15]

Data sources would include wearable devices (e.g., fitness trackers, smartwatches, and ECG monitors), sensors, health applications (e.g., EHRs), medical imaging systems, smart home technologies, and laboratory reports covering domains such as radiology, pathology, dermatology, mammography, and genomics.

SPARK would capture various physiological data in real time, including heart rate, sleep patterns, glucose levels, body temperature, oxygen saturation, respiratory rate, ECG signals, blood pressure, and activity metrics such as steps, distance, and calories burned.[4,13,15,23]

Data Preprocessing and Storage:

The cloud-based backend would manage the core logic of the system, control data flow, and ensure efficient communication between external entities and internal components in accordance with the programmed logic.

A cloud-based database would store seniors' basic information (name, age, weight, height, and BMI); health history (e.g., surgical and medication records); existing conditions; health tips; personal goals; advised tasks; current medications; and raw and preprocessed health data with timestamps and source identifiers for traceability and auditing, along with system-generated reports, allowing access to the data whenever needed.

The backend would communicate with APIs of external data sources to aggregate health data and also collect prompts and data entered through the UI. It would securely store them in the database via the database's API. The backend would then preprocess the data by cleaning and structuring it and store this version in the database as well. It would structure user prompts into a format compatible with the FM/LLM.

It would de-identify sensitive information to ensure compliance with privacy, security, and regulatory standards, such as the Health Insurance Portability and Accountability Act (HIPAA), and provide this data, along with the structured user prompts and other relevant information retrieved from the database, to the FM/LLM via the model's API.[4,18,27]

Data Processing and Output Generation:

A biomedical FM/LLM with generalization capabilities across multiple tasks, pre-trained on vast and diverse datasets, and adaptable to new tasks or data types through fine-tuning or prompt engineering, would be leveraged by in the system.

The model would accept input multimodal health data and prompts, analyze and interpret them accurately, and generate reliable reports in response to prompts for risk prediction, disease diagnosis, treatment recommendations, or as summaries based on user settings.

The backend would collect these reports and store them in the database.

Output Presentation and Notifications:

The backend would display the output responses and reports on the user's dashboard or provide them through voice-based interaction, as requested. It would also send notifications, such as reminders and alerts, to seniors, care teams, or emergency services based on user preferences and permissions.

Evaluation and Refinement:

SPARK would be evaluated through a structured evaluation plan that combines controlled environment testing with standardized measurements. A human-in-the-loop approach would be incorporated throughout the evaluation process. Access controls, multiple layers of data protection, data de-identification, and compliance with privacy and regulatory standards such as HIPAA would be implemented to safeguard all information. In the controlled testing environment, human verifiers or synthetic patient data would provide inputs via the UI and relevant health sources. SPARK would incorporate a feedback mechanism in which human verifiers, including medical providers, could assess the system's usability, fairness, and accuracy of outputs.[14] The insights from these evaluations would be used to refine SPARK iteratively. Testers would assess SPARK on metrics such as reliability, robustness, and stability.[14] It would also be evaluated for interoperability across all internal and external components, response time, and overall technical performance.

Additionally, the integrated FM/LLM within SPARK would be evaluated using benchmarks such as HealthBench, MultiMedBench, DRAGON, and MultiMedQA; datasets such as MedQA, SLAKE, and PathVQA; and standardized metrics such as F1-RadGraph and Accuracy.[17,28,30]

System Deployment and Maintenance:

SPARK would be deployed as a scalable, cloud-based platform for real-world use, accessible through web portals and mobile applications, offering anytime, anywhere access for seniors, family members, and healthcare providers. It would ensure transparency and maintain documentation to support explainability and reliability.

Maintenance would involve continuous monitoring for risks and adverse events, feedback-based refinements, regular system audits (e.g., Responsible AI (RAI) black box assessments and co-versioning registry), and routine software updates to maintain compliance, accuracy, and efficiency.[14,18] These measures would ensure SPARK would remain a reliable, high-quality platform for all stakeholders.

Challenges

Ensuring the effectiveness, usability, accuracy, and ethical deployment of the model and the system remains a central challenge. Key challenges include:

- Model accessibility: Open-source models, such as BiomedGPT, have significant potential for integration with SPARK. In contrast, closed-source models, like Med-PaLM M, cannot be used directly but can serve as references for identifying similar models.
- Model selection: SPARK requires evaluation of multimodal, multitask biomedical FMs and LLMs that can provide risk predictions, disease diagnoses, treatment recommendations, and summaries. Selection focuses on models trained on governed data, designed with ethical principles, and subjected to ethical audits to ensure transparency, explainability, accuracy, fairness, and safety.[5,18]
- Bias mitigation and fairness: Model outputs may reflect bias due to training on large-scale, unstructured online data.[12] SPARK aims to leverage a carefully developed FM/LLM and incorporate human feedback and ethical audits to promote fairness and generate unbiased outputs.[8]
- Hallucinations and output reliability: Models may generate outputs that deviate from user input, context, or facts. SPARK would interact with FM/LLM trained using techniques such as RLHF and incorporate human verification and safety prompts to improve output accuracy and reliability.[5,12]
- Privacy and data protection: SPARK would comply with regulatory standards, such as HIPAA, and implement strict access controls, data de-identification, and multiple layers of data protection for sensitive information, including seniors' social demographics and medical histories.[13]

DOI: 10.36838/IJHSR89.193

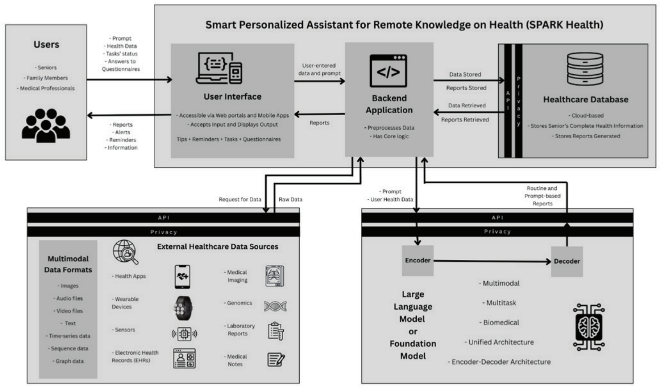

Figure 4: System Architecture of SPARK Health. The diagram depicts that SPARK Health would comprise User Interface, Backend, and Database, and would interact with Users, Healthcare Data Sources, and FM/LLM.[4,18,27]

Conclusion

This paper systematically reviewed AI technologies in healthcare, primarily FMs and LLMs, highlighting their potential to support healthcare. Based on this review, leveraging a reliable, generative, multimodal, multitask biomedical FM/LLM, potentially built on a unified, encoder-decoder transformer architecture, represents a promising approach for virtual real-time health monitoring and healthcare decision-making. The paper also introduced a conceptual proposal for SPARK Health, a family-managed, personalized healthcare AI assistant, along with its system architecture and potential applications for remote care of seniors. SPARK Health would be implemented; an FM/LLM of the identified type would be evaluated and leveraged; and the complete setup would be assessed, deployed, and maintained. Given the projected shortage of 18 million healthcare professionals by 2030, based on current trends in global healthcare demand, AI systems have the potential to address this shortage effectively.[14] SPARK Health aims to address the urgent challenges of remote senior care and promote proactive healthcare via an integrated platform accessible to multiple stakeholders. It is designed to help seniors live their later years in good health, comfort, and dignity, while enabling families and healthcare providers to manage care efficiently. SPARK Health is conceptualized as an innovative solution that integrates with digital health platforms and leverages a multimodal, multitask biomedical FM/LLM to provide comprehensive, personalized, proactive, conversational, and collaborative real-time remote healthcare support to seniors, managed by families with healthcare provider involvement. With its primary goal of supporting the well-being of seniors, their families, medical providers, and healthcare systems, SPARK Health aims to align with the UN Decade of Healthy Ageing (2021-2030).[3] Technology's ultimate purpose is to benefit humanity, and SPARK Health, as a human-AI technological solution, is envisioned to pursue that purpose.

Acknowledgments

I sincerely thank Dr. Siddharth Krishnan, Mr. Samuel Lefcourt, and Mr. Niraj Ganesh for their guidance, encouragement, and review of this paper. I also extend my gratitude to the seniors and families who inspired this research.

References

1. United Nations (UN). World Social Report 2023: Leaving No One Behind in an Ageing World. https://social.desa.un.org/sites/default/files/publications/2023-02/WorldSocialReport2023.pdf
2. World Health Organization (WHO). Promoting Healthy Ageing in a Digital World. 2024. https://www.decadeofhealthyageing.org/docs/librariesprovider5/default-document-library/terms-of-reference---healthy-ageing-collaborative.pdf
3. World Health Organization (WHO). Healthy Ageing Collaborative. https://www.decadeofhealthyageing.org/about/join-us/collaborative
4. Nigar, N. AI in Remote Patient Monitoring. arXiv 2024. https://doi.org/10.48550/arXiv.2407.17494
5. Alkaeed, M.; Abioye, S.; Qayyum, A.; Mekki, Y. M.; Berrou, I.; Abdallah, M.; Al-Fuqaha, A.; Bilal, M.; Qadir, J. Open Foundation Models in Healthcare: Challenges, Paradoxes, and Opportunities with GenAI Driven Personalized Prescription. arXiv 2025. https://doi.org/10.48550/arXiv.2502.04356
6. Arora, R. K.; Wei, J.; Hicks, R. S.; Bowman, P.; Quiñonero-Candela, J.; Tsimpourlas, F.; Sharman, M.; Shah, M.; Vallone, A.; Beutel, A.; Heidecke, J.; Singhal, K. HealthBench: Evaluating Large Language Models Towards Improved Human Health. arXiv 2025. https://doi.org/10.48550/arXiv.2505.08775
7. Mello-Thoms, C.; Mello, C. A. B. Clinical Applications of Artificial Intelligence in Radiology. The British Journal of Radiology 2023, 96 (1150), 20221031. https://doi.org/10.1259/bjr.20221031
8. Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E.; Brynjolfsson, E.; Buch, S.; Card, D.; Castellon, R.; Chatterji, N.; Chen, A.; Creel, K.; Davis, J. Q.; Demszky, D.; Donahue, C.; Doumbouya, M.; Durmus, E.; Ermon, S.; Etchemendy, J.; Ethayarajh, K.; Fei-Fei, L.; Finn, C.; Gale, T.; Gillespie, L.; Goel, K.; Goodman, N.; Grossman, S.; Guha, N.; Hashimoto, T.; Henderson, P.; Hewitt, J.; Ho, D. E.; Hong, J.; Hsu, K.; Huang, J.; Icard, T.; Jain, S.; Jurafsky, D.; Kalluri, P.; Karamcheti, S.; Keeling, G.; Khani, F.; Khattab, O.; Koh, P. W.; Krass, M.; Krishna, R.; Kuditipudi, R.; Kumar, A.; Ladhak, F.; Lee, M.; Lee, T.; Leskovec, J.; Levent, I.; Li, X. L.; Li, X.; Ma, T.; Malik, A.; Manning, C. D.; Mirchandani, S.; Mitchell, E.; Munyikwa, Z.; Nair, S.; Narayan, A.; Narayanan, D.; Newman, B.; Nie, A.; Niebles, J. C.; Nilforoshan, H.; Nyarko, J.; Ogut, G.; Orr, L.; Papadimitriou, I.; Park, J. S.; Piech, C.; Portelance, E.; Potts, C.; Raghunathan, A.; Reich, R.; Ren, H.; Rong, F.; Roohani, Y.; Ruiz, C.; Ryan, J.; Ré, C.; Sadigh, D.; Sagawa, S.; Santhanam, K.; Shih, A.; Srinivasan, K.; Tamkin, A.; Taori, R.; Thomas, A. W.; Tramèr, F.; Wang, R. E.; Wang, W.; Wu, B.; Wu, J.; Wu, Y.; Xie, S. M.; Yasunaga, M.; You, J.; Zaharia, M.; Zhang, M.; Zhang, T.; Zhang, X.; Zhang, Y.; Zheng, L.; Zhou, K.; Liang, P. On the Opportunities and Risks of Foundation Models. arXiv 2021. https://doi.org/10.48550/arXiv.2108.07258
9. Capogrosso, L.; Cunico, F.; Cheng, D. S.; Fummi, F.; Cristani, M. A Machine Learning-Oriented Survey on Tiny Machine Learning. arXiv 2023. https://doi.org/10.48550/arXiv.2309.11932
10. Hagos, D. H.; Battle, R.; Rawat, D. B. Recent Advances in Generative AI and Large Language Models: Current Status, Challenges, and Perspectives. arXiv 2024. https://doi.org/10.48550/arXiv.2407.14962
11. Wang, Z.; Wu, Z.; Agarwal, D.; Sun, J. MedCLIP: Contrastive Learning from Unpaired Medical Images and Text. arXiv 2022. https://doi.org/10.48550/arXiv.2210.10163

12. Wang, Z.; Chu, Z.; Doan, T. V.; Ni, S.; Yang, M.; Zhang, W. History, Development, and Principles of Large Language Models-An Introductory Survey. arXiv 2024. https://doi.org/10.48550/arXiv.2402.06853
13. Li, L.; Zhou, J.; Gao, Z.; Hua, W.; Fan, L.; Yu, H.; Hagen, L.; Zhang, Y.; Assimes, T. L.; Hemphill, L.; Ma, S. A Scoping Review of Using Large Language Models (LLMs) to Investigate Electronic Health Records (EHRs). arXiv 2024. https://doi.org/10.48550/arXiv.2405.03066
14. Bajwa, J.; Munir, U.; Nori, A.; Williams, B. Artificial Intelligence in Healthcare: Transforming the Practice of Medicine. Future Healthcare Journal 2021, 8 (2), e188–e194. https://doi.org/10.7861/fhj.2021-0095
15. Tu, T.; Azizi, S.; Driess, D.; Schaekermann, M.; Amin, M.; Chang, P.-C.; Carroll, A.; Lau, C.; Tanno, R.; Ktena, I.; Mustafa, B.; Chowdhery, A.; Liu, Y.; Kornblith, S.; Fleet, D.; Mansfield, P.; Prakash, S.; Wong, R.; Virmani, S.; Semturs, C.; Mahdavi, S. S.; Green, B.; Dominowska, E.; Arcas, B. A. y; Barral, J.; Webster, D.; Corrado, G. S.; Matias, Y.; Singhal, K.; Florence, P.; Karthikesalingam, A.; Natarajan, V. Towards Generalist Biomedical AI. arXiv 2023. https://doi.org/10.48550/arXiv.2307.14334
16. Huo, C.; Chen, K.; Zhang, S.; Wang, Z.; Yan, H.; Shen, J.; Hong, Y.; Qi, G.; Fang, H.; Wang, Z. When Remote Sensing Meets Foundation Model: A Survey and Beyond. Remote Sensing 2025, 17 (2), 179. https://doi.org/10.3390/rs17020179
17. Yuan, M.; Bao, P.; Yuan, J.; Shen, Y.; Chen, Z.; Xie, Y.; Zhao, J.; Chen, Y.; Zhang, L.; Shen, L.; Dong, B. Large Language Models Illuminate a Progressive Pathway to Artificial Healthcare Assistant: A Review. arXiv 2023. https://doi.org/10.48550/arXiv.2311.01918
18. Lu, Q.; Zhu, L.; Xu, X.; Liu, Y.; Xing, Z.; Whittle, J. A Taxonomy of Foundation Model Based Systems through the Lens of Software Architecture. arXiv 2023. https://doi.org/10.48550/arXiv.2305.05352
19. Maity, S.; Saikia, M. J. Large Language Models in Healthcare and Medical Applications: A Review. Bioengineering 2025, 12 (6), 631. https://doi.org/10.3390/bioengineering12060631
20. Li, Y.; Li, Z.; Zhang, K.; Dan, R.; Jiang, S.; Zhang, Y. ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge. Cureus 2023. https://doi.org/10.7759/cureus.40895
21. Gupta, A.; Sawhney, S.; Ahmed, S. CareTaker.ai—A Smart Health-Monitoring and Caretaker-Assistant System for Elder Healthcare. In The 1st International Conference on AI Sensors & the 10th International Symposium on Sensor Science; MDPI, 2025; p 7. https://doi.org/10.3390/engproc2024078007
22. Cosentino, J.; Belyaeva, A.; Liu, X.; Furlotte, N. A.; Yang, Z.; Lee, C.; Schenck, E.; Patel, Y.; Cui, J.; Schneider, L. D.; Bryant, R.; Gomes, R. G.; Jiang, A.; Lee, R.; Liu, Y.; Perez, J.; Rogers, J. K.; Speed, C.; Tailor, S.; Walker, M.; Yu, J.; Althoff, T.; Heneghan, C.; Hernandez, J.; Malhotra, M.; Stern, L.; Matias, Y.; Corrado, G. S.; Patel, S.; Shetty, S.; Zhan, J.; Prabhakara, S.; McDuff, D.; McLean, C. Y. Towards a Personal Health Large Language Model. arXiv 2024. https://doi.org/10.48550/arXiv.2406.06474
23. AlSaad, R.; Abd-alrazaq, A.; Boughorbel, S.; Ahmed, A.; Renault, M.-A.; Damseh, R.; Sheikh, J. Multimodal Large Language Models in Health Care: Applications, Challenges, and Future Outlook. J Med Internet Res 2024, 26, e59505. https://doi.org/10.2196/59505
24. Wang, T.; Roberts, A.; Hesslow, D.; Scao, T. L.; Chung, H. W.; Beltagy, I.; Launay, J.; Raffel, C. What Language Model Architecture and Pretraining Objective Work Best for Zero-Shot Generalization? arXiv 2022. https://doi.org/10.48550/arXiv.2204.05832
25. Shaik, T.; Tao, X.; Higgins, N.; Li, L.; Gururajan, R.; Zhou, X.; Acharya, U. R. Remote Patient Monitoring Using Artificial Intelligence: Current State, Applications, and Challenges. WIREs Data Mining & Knowledge Discovery. 2023, 13 (2), e1485. https://doi.org/10.1002/widm.1485
26. Yang, B.; Jiang, S.; Xu, L.; Liu, K.; Li, H.; Xing, G.; Chen, H.; Jiang, X.; Yan, Z. DrHouse: An LLM-Empowered Diagnostic Reasoning System through Harnessing Outcomes from Sensor Data and Expert Knowledge. arXiv 2025. https://doi.org/10.48550/arXiv.2405.12541
27. Abbasian, M.; Azimi, I.; Rahmani, A. M.; Jain, R. Conversational Health Agents: A Personalized LLM-Powered Agent Framework. arXiv 2023. https://doi.org/10.48550/arXiv.2310.02374
28. Zhang, K.; Zhou, R.; Adhikarla, E.; Yan, Z.; Liu, Y.; Yu, J.; Liu, Z.; Chen, X.; Davison, B. D.; Ren, H.; Huang, J.; Chen, C.; Zhou, Y.; Fu, S.; Liu, W.; Liu, T.; Li, X.; Chen, Y.; He, L.; Zou, J.; Li, Q.; Liu, H.; Sun, L. BiomedGPT: A Generalist Vision-Language Foundation Model for Diverse Biomedical Tasks. arXiv 2023. https://doi.org/10.48550/arXiv.2305.17100
29. Builtjes, L.; Bosma, J.; Prokop, M.; van Ginneken, B.; Hering, A. Leveraging Open-Source Large Language Models for Clinical Information Extraction in Resource-Constrained Settings. arXiv 2025. https://doi.org/10.48550/arXiv.2507.20859
30. Kim, Y.; Xu, X.; McDuff, D.; Breazeal, C.; Park, H. W. Health-LLM: Large Language Models for Health Prediction via Wearable Sensor Data. arXiv 2024. https://doi.org/10.48550/arXiv.2401.06866

■ Author

Karthik Reddy is a senior at Temple City High School in California. He is an incoming Electrical Engineering and Computer Sciences student at the University of California, Berkeley, focused on developing innovative technological systems for academic and global impact.