

# Generative AI Healthcare Agents

## Team 3

- Aparna Bharathi Suresh
- Pavan Srivatsav Devarakonda
- Sai Naga Sanjana Chippada
- Shravani Dattaram Gawade
- Sujata Deepraj Joshi

# Project Background

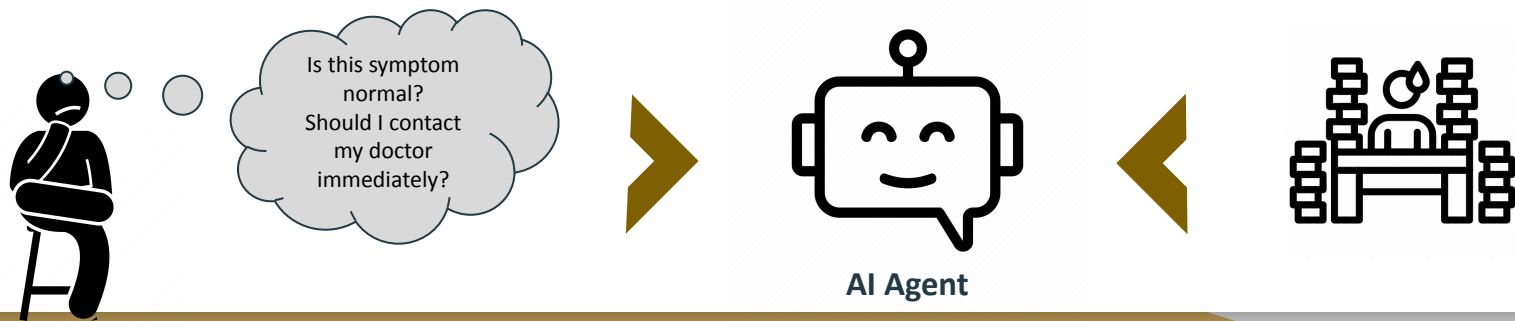
**Motivation** - Chronic conditions like cardiac, respiratory, and neurological disorders require continuous care. However, resource constraints in healthcare systems delay timely assistance and increase professional workload.

**Need**- Reduce healthcare professional workload & improve timely support for patients through AI-powered tools.

**Our Solution** - Developing **Generative AI Healthcare Agents** using a **Retrieval-Augmented Generation (RAG)** framework integrated with specialized **Large Language Models (LLMs)**.

## Impact

- Enhance continuous patient engagement
- Provide evidence-based medical insights
- Improve access to early guidance and support for chronic diseases



# Executive Summary

## Key Components -

**Medical Information Retrieval Agent** - Retrieves real-time, evidence-based medical information from PubMed, and other reputable sources to answer queries about conditions, symptoms, and treatments.

**Symptom Checker Agent** - Evaluates patient symptoms to suggest potential conditions or recommend medical attention, crucial for overlapping symptoms in chronic diseases.

## Technology Stack -

- **LLMs:** BioMistral-7B, MedLlama3, BioGpt, MediTron
- **Training Data:** MIMIC-III, PubMedQA, MedDialog
- **Deployment:** Google Cloud Platform (GCP)

**Goal** - Deliver an accessible, real-time & reliable AI-driven healthcare support for chronic disease patients.

# Project Requirements

## Technical Requirements -

- Models like BioMistral-7B, MedLlama3, BioGpt, and MediTron will be fine-tuned, with performance evaluated by F1 score, precision, and recall on datasets like MedQA, PubMedQA, MIMIC\_III and MIMIC\_IV.

## Functional Requirements -

- Accurate Medical Responses: Evaluated using BLEU, ROUGE, METEOR, FCS, GEval, LLM Judge scores.
- Real-Time Information Retrieval: Implemented via RAG (Retrieval-Augmented Generation) framework.
- Medical Concept Recall (MCR): Ensure high recall of key clinical concepts.

## Performance Requirements -

- Multi-User Support: System handles concurrent users with acceptable latency and throughput.

# Project Requirements

## **Datasets**

- Fine-tuning will use MIMIC-III, MIMIC-IV, PubMed, and MedDialog datasets

## **Pinecone Vector Database**

- Preprocessed medical knowledge will be stored in a Pinecone vector database for efficient retrieval

# Team Organization

## Team Members-

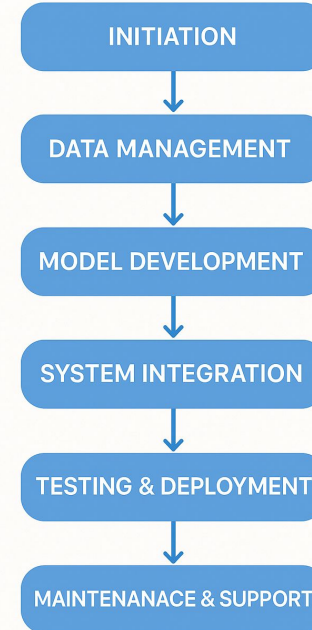
Aparna Suresh, Pavan Srivatsav Devarakonda, Sai Naga Sanjana Chippada, Shravani Gawade, Sujata Joshi

**Faculty Advisor:** Dr. Simon Shim, SJSU

## Organization Model - Structured into 6 Phases

Each phase has clear **milestones**, **deliverables**, and **work packages (WPs)**.

## ORGANIZATION MODEL



# Function Roles

Team Member	Roles	Responsibilities
Aparna S	Backend Developer	GCP backend deployment, API development
Sanjana C	Frontend Developer	Chatbot UI with Streamlit and API integration
Pavan D	ML Engineer	Fine-tuning LLMs, RAG integration, model evaluation
Shravani G	Deployment Engineer	Orchestrate CI/CD pipelines for model training, automate ELT
Sujata J	Data Engineer	ELT pipeline, GCP bucket

# Literature Survey

Paper / Source	Focus Area	Techniques Used	Findings	Limitations / Gaps
<b>Sandeep Reddy</b>	Implementation of GenAI in Healthcare	TAM, NASSS Frameworks	Defines translational path for safe integration	Operational scalability and governance need work
<b>Hiesinger et al. (Almanac)</b>	Retrieval-Augmented Models for Clinical QA	RAG, Evaluation by physicians	Improves factual accuracy and safety over base LLMs	Model reliability still requires medical supervision
<b>Akhil Vaid et al.</b>	Autonomous Clinical Agents	RAG, LLMs, Simulated Hospital Scenarios	Emulates clinician decision-making	Performance varies by specialty; lacks error resilience
<b>Kiran Gulia et al.</b>	ML-Powered Digital Twins (PDTs)	ML, Personalization, Ethical AI	Improves diabetes treatment via individualized prediction	Ethical implications, data privacy concerns
<b>Chen et al.</b>	Democratization of GenAI	Narrative review of GenAI in MedEd & Clinical Use	GenAI can bridge health equity gaps	Bias, safe deployment in underserved populations



# Literature Survey

Paper / Source	Focus Area	Techniques Used	Findings	Limitations / Gaps
<b>Milne-Ives et al.</b> (Systematic Review)	Conversational Agents in Healthcare	NLP, Chatbots	High user satisfaction and usability	Lack of robust evaluation, low interactivity, cost concerns
<b>Sulis et al.</b>	Agent-Based Healthcare Systems	Multi-agent systems (MAS), ABM	Agents improve hospital workflow and patient engagement	Trust, interoperability, integration with ML needed
<b>Narayan et al.</b>	AI Agents for Chronic Conditions	ML, Dialog Systems	Conversational agents effective in chronic care	Evaluation methods inconsistent, long-term impact unclear
<b>Chetty et al.</b>	Generative AI in Diagnosis & Treatment	GANs, LLMs	Enhances diagnostics, patient interaction, drug discovery	Ethics, privacy, transparency challenges
<b>Siva Sai et al.</b>	GAI Use Cases in Clinical Settings	ChatGPT, DALL·E	Supports medical imaging, clinical trials, patient education	Bias, model accuracy, regulation enforcement

# Key findings of Literature Survey

These findings emphasize the critical need for:-

- **Domain-Specific LLMs with RAG:** Ensures accurate, context-aware, and real-time medical responses.
- **Ethically Governed AI Systems:** Frameworks that prioritize data privacy, explainability, and fairness.
- **Standardized Evaluation Protocols:** Metrics like **Factual Consistency Score (FCS)**, **MCR**, **ROUGE**, **BERTScore** for consistent benchmarking.
- **Real-World Piloting & Feedback Loops:** Integration into hospital settings to gather clinician and patient feedback.
- **Equity-Focused AI Development:** Ensuring models are inclusive and adaptable to underserved populations.

# Project Resources, Technology and Platform

## Hardware

Component	Specifications	Justification
GCP Cloud Instances	NVIDIA A100 GPU, 128GB RAM, multi-core CPUs	Required for LLM training, inference, and real-time responses

## Software

Tool	Purpose	Justification
Python, PyTorch, Pandas, Scikit-learn, TensorFlow	Model training, fine-tuning	Industry-standard for deep learning workflows
Vertex AI (GCP)	End-to-end model training, evaluation, deployment	Simplifies model lifecycle, integrates with GCP infrastructure
Streamlit	Chatbot UI Development	Enables rapid development of patient-friendly interfaces
Pinecone	Vector database for embedding search	Powers fast semantic retrieval in RAG pipeline
Cloud Composer / Airflow	ELT orchestration	Automates data preprocessing, transformation, and scheduling

# Project Resources, Technology and Platform

## Security & Compliance Tools

Tool	Purpose	Justification
GCP IAM	Role-based access control (RBAC)	Enforces HIPAA compliance and auditability
AES-256 Encryption	Encrypts data at rest and in transit	Required for secure medical data handling
Audit Logs	Monitor access and changes	Ensures accountability and data integrity

# Technology Survey

Component	Purpose	Tools / Technology
<b>LLMs</b>	Understand and generate medical text	BioMistral-7B, MedLlama3, BioGPT, Meditron
<b>RAG Framework</b>	Real-time, evidence-based responses	JMLR (Joint Medical LLM + Retrieval), Pinecone
<b>Vector Database</b>	Store/retrieve embeddings for semantic search	Pinecone
<b>Cloud Infrastructure</b>	Scalable, secure deployment	Google Cloud Platform (GCP)
<b>User Interface</b>	Patient-facing chatbot	Streamlit, HTML/CSS
<b>Security &amp; Compliance</b>	Data privacy, HIPAA compliance	GCP IAM, AES-256 encryption
<b>Model Fine-Tuning</b>	Specialize LLMs for healthcare	PyTorch, LoRA, Hugging Face
<b>Model Evaluation</b>	Ensure response accuracy and safety	BERTScore, BLEU, ROUGE, MCR, Factual Consistency Score

# Big Data Statistics and Datasets

## MIMIC-IV

- De-identified ICU patient records: 7,195,726 rows (CSV, ~717MB)
- Includes vitals, diagnoses, medications, and triage data
- Used for clinical reasoning and patient scenario modeling

## MIMIC-III

- Structured ICU dataset containing ~59,000 hospital admissions and over 2 million clinical notes.
- Used primarily for **question-answer (QA) generation**, derived from real clinical cases
- Project utilized **1,287 QA pairs** from test\_final.json in **JSON format** (~a few MB)

# Big Data Statistics and Datasets

## iCliniq

- 7,320 rows (~9MB)
- Real-world patient-doctor QA pairs from Hugging Face
- Used for conversational fine-tuning

## MedQuAD

- 47,457 rows (~22MB)
- Curated medical QA from trusted health sources (MedlinePlus, Cancer.gov)
- Includes "focus\_area" for disease-topic labeling

# Big Data Statistics and Datasets

## Processed Data Shape & Structure

- Final merged dataset: **7.1 million rows × 16 columns**
- Merged across subject\_id and stay\_id fields using outer joins
- Redundant and irrelevant fields (e.g., intime, race) were dropped

## Preprocessing Techniques

- Missing value imputation, standardization, and deduplication
- ICD-9 & ICD-10 code mapping for diagnosis enrichment
- Preprocessed Mimic3, Mimic4, MedQuad, and iCliniq Q/A datasets, standardizing them into Instruction (query), Input (patient details), and Output (answer) format, ensuring compatibility for LLM training and fine-tuning to generate precise, contextually relevant answers.



# ML Model Selection & Innovation

We have chosen Healthcare-Specific Language Models

## Selected LLMs

- MedLlama3 – Built on LLaMA-2/3; handles general health Q&A and patient education.
- BioMistral-7B – PubMed-trained; excels in biomedical and clinical literature retrieval.
- BioGPT – GPT-style model; optimized for medical text generation and patient-facing responses.
- Meditron – High medical reasoning power; supports diagnostic assistance and complex queries.

## Model Innovation

- **RAG + JMLR** - Retrieval-Augmented Generation enhanced with Joint Medical LLM and Retrieval Training to reduce hallucination.
- **Vector Database (Pinecone)** - Stores real-time vectorized PubMed data for fast similarity-based retrieval.
- **Optimization Techniques** - LoRA fine-tuning, instruction tuning (BioGPT-JSL), and quantization for deployment efficiency.

# Comparison & Justification


Model	Primary Use Case	Key Strengths	Justification
<b>MedLlama3</b>	General healthcare Q&A, triage	High versatility, accurate contextual reasoning	Serves as core model for broad patient queries and education; fine-tuned on iCliniq & MedDialog
<b>BioMistral-7B</b>	Biomedical info retrieval	Strong clinical terminology, multilingual support	Specialized for PubMed-style queries; fine-tuned on PubMedQA; BERTScore $\uparrow$ 0.82 $\rightarrow$ 0.92
<b>BioGPT</b>	Patient education, conversational Q&A	Fluent generative text, lightweight deployment	Ideal for open-ended patient queries; supports instruction tuning (BioGPT-JSL)
<b>Meditron</b>	Diagnostic reasoning, decision support	High MedQA accuracy (77.6%), multimodal capabilities (Meditron-V)	Excels in complex medical logic; trained on GAP-Replay corpus (48.1B tokens)

# Machine Learning Prototyping

- All models fine-tuned using **LoRA** (parameter-efficient training)
- Trained on cleaned and preprocessed medical datasets:
  - MIMIC-III, MIMIC-IV (ICU records)
  - MedDialog, iCliniq (doctor-patient Q&A)
  - PubMedQA (medical QA benchmark)
- Training and evaluation on **Google Cloud Vertex AI** using A100 GPUs
- Stored medical knowledge embeddings in **Pinecone** for RAG integration
- QA data processed to JSONL format with fields: instruction, input, output

# Results from Previous Demos

# Data Processing through DAGs



[DAGs](#)
[Cluster Activity](#)
[Datasets](#)
[Browse](#)
[Admin](#)
[Docs](#)
[Composer](#)
22:07 UTC

## project298

All 3
Active 3
Paused 0
Running 0
Failed 0


Auto-refresh

DAG	Owner	Runs	Schedule	Last Run	Next Run	Recent Tasks	Actions	Links
<input checked="" type="checkbox"/> airflow_monitoring	airflow	<div> <div>1</div> <div>2</div> <div>3</div> <div>4</div> </div>	*15 * * * *	2025-04-09, 20:00:00	2025-04-09, 20:00:00	<div> <div>1</div> <div>2</div> <div>3</div> <div>4</div> <div>5</div> <div>6</div> <div>7</div> <div>8</div> <div>9</div> <div>10</div> </div>	<a href="#">▶</a> <a href="#">🔍</a> <a href="#">🛑</a>	...
<input type="checkbox"/> mimic_merge_pipeline <span>gcs merge mimic</span>	airflow	<div> <div>1</div> <div>2</div> <div>3</div> <div>4</div> </div>	@daily	2025-04-09, 09:36:12	2025-04-09, 00:00:00	<div> <div>1</div> <div>2</div> <div>3</div> <div>4</div> <div>5</div> <div>6</div> <div>7</div> <div>8</div> <div>9</div> <div>10</div> </div>	<a href="#">▶</a> <a href="#">🔍</a> <a href="#">🛑</a>	...
<input type="checkbox"/> mimic_preprocessing_pipeline <span>gcs mimic preprocessing</span>	airflow	<div> <div>1</div> <div>2</div> <div>3</div> <div>4</div> </div>	@daily	2025-04-09, 18:49:17	2025-04-09, 00:00:00	<div> <div>1</div> <div>2</div> <div>3</div> <div>4</div> <div>5</div> <div>6</div> <div>7</div> <div>8</div> <div>9</div> <div>10</div> </div>	<a href="#">▶</a> <a href="#">🔍</a> <a href="#">🛑</a>	...

1
2
3

Showing 1-3 of 3 DAGs

# Data Transformation

Cloud Storage

Overview

Buckets

Monitoring

Settings

Storage Intelligence

Insights datasets

Configuration

Bucket details

Go to path

Refresh

Learn

Objects

Configuration

Permissions

Protection

Lifecycle

Observability

New

Inventory Reports

Operations

Folder browser

data298a

final/

intermediate/

processed/

Buckets > data298a > final

Create folder

Upload

Transfer data

Other services


Filter by name prefix only

Filter

Filter objects and folders

Show

Live objects only

<input type="checkbox"/>	Name	Size	Type	Created	Storage class	Last r
<input type="checkbox"/>	 merged_final.csv	2.1 GB	text/csv	Apr 9, 2025, 2:40:29 AM	Standard	Apr 9, <div></div>

# Evaluation Metrics Comparison Table

Metric	Baseline (pre-trained)	Fine-Tuned (Healthcare-Specific)
BERTScore Precision	0.8	0.8
BERTScore Recall	0.86	0.89
BERTScore F1	0.83	0.84
ROUGE-L	0.08	0.08
Entity F1	0.2	0.1
FCS (Factual Consistency Score)	0.3859	0.53
MCR (Medical Concept Recall)	0.2	0.2
BLEU	0.01	0.02
Hybrid BERT-BLEU	0.58	0.59
LLM Judge Score	0.52	0.84
GEval Score	0.44	0.75
METEOR	0.16	0.24



Thank You