MediAssist: LLM-Powered Healthcare Intelligence

Personalized Patient Engagement
&
Medicaid Policy Navigator

Business Applications for LLMs

Presenters:
Shashwat Kumar (sk5520)
Somit Jain (sj3396)

Introduction

- Developed **two RAG-based systems** — *Patient Engagement Assistant* & *Medicaid Policy Navigator*.
- Tackles **low health literacy** and **unsearchable Medicaid documentation**.
- Leverages **open-source LLMs** (*Qwen-2.5*, *BGE*, *MarianMT*) for **efficient, transparent, and reproducible workflows**.
- Converts **complex medical guidance** into **bilingual, patient-friendly discharge instructions**.
- Transforms **policy PDFs** into **searchable, citation-grounded knowledge bases**.
- Runs **entirely on free Google Colab GPUs**, using **modular and reproducible pipelines**.
- Enhances **accessibility, trust, and real-time retrieval** in healthcare communication.

Dataset

OpenFDA:

- A public FDA database providing structured data on drugs, devices, and adverse events for transparency and safety analysis.

MedLinePlus:

- A consumer-facing NIH database offering easy-to-understand, verified medical and drug information.

MedicAid Policy Documents:

- Repository of **state-issued policy bulletins** detailing coverage, billing, and reimbursement updates across pharmacy, dental, transportation, and managed-care programs.

- Each document includes **official guidance on benefits, provider requirements, and regulatory changes**, forming the basis for a **retrieval-based policy navigator** that enables searchable, citation-linked insights.

Patient Portal - Problem Statement and Proposed Architecture

Problem Statement

- Existing drug information platforms (e.g., **OpenFDA**, **MedlinePlus**) are **not patient-tailored**.
- Discharge instructions are **static, English-only**, and **not optimized for readability**.
- Generating **customized, multilingual discharge summaries** still requires **manual effort**.

Proposed Architecture

- **Data Ingestion** → Retrieve verified medical content from *MedlinePlus* and *OpenFDA*.
- **Chunking & Annotation** → Segment text into coherent sections and attach metadata.
- **Embedding & Indexing** → Generate dense vector representations for efficient semantic retrieval.
- **Grounded Generation** → Produce factual, section-based summaries using a compact LLM constrained by retrieved context.
- **Accessibility Layer** → Evaluate readability, translate to Spanish, and synthesize audio narration for low-literacy users.

Patient Portal - Workflow

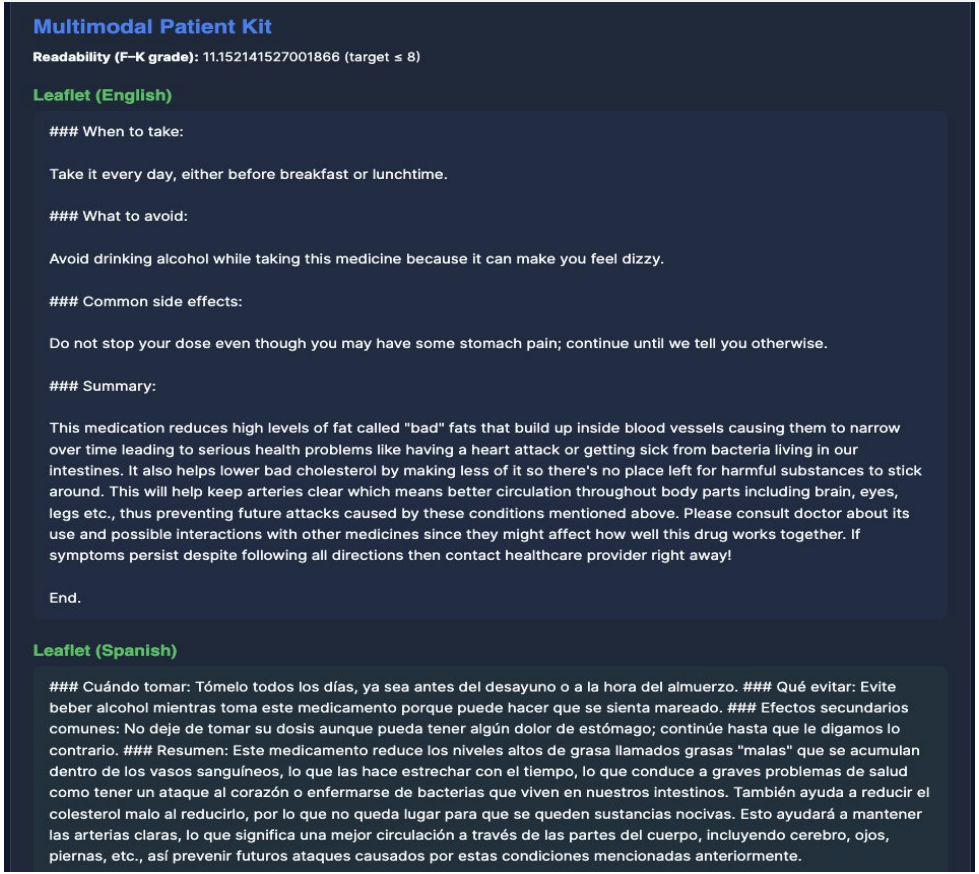
- **Patient Profiling** → Specify **age**, **literacy level**, **language**, and **target drug** to enable personalized instruction generation.
- **Data Retrieval** → Fetch **verified medical content** from *MedlinePlus* and *OpenFDA* APIs.
- **Preprocessing & Chunking** → Clean, normalize, and segment documents into **metadata-tagged sections** for context alignment.
- **Embedding & Indexing** → Generate **dense semantic vectors** using *BGE-small* and store them in a **FAISS index** for efficient retrieval.
- **LLM-Guided Generation** → Use *Qwen 2.5 (1.5B)* to produce **factual, section-based discharge instructions** grounded in retrieved context.
- **Readability & Accessibility** → Simplify language, **translate to Spanish** via *MarianMT*, and **generate audio narration** using *gTTS* for low-literacy users.
- **Final Output** → Deliver a **bilingual, audio-enabled medical leaflet** that is **personalized, verifiable, and patient-friendly**.

Patient Portal - Technical Architecture

Layer	Function	Libraries / Models
Data Acquisition	Fetch structured/unstructured data from openFDA & MedlinePlus	requests, BeautifulSoup
Preprocessing	Clean, segment, and annotate data	regex, pandas, numpy
EDA / Visualization	TF-IDF term ranking and word cloud for validation	scikit-learn, matplotlib, wordcloud
Retrieval Layer	Semantic search over chunked text	sentence-transformers, FAISS
Generation Layer	Context-aware summarization	Qwen-2.5-1.5B-Instruct, TinyLlama
Accessibility Layer	Readability, translation, TTS output	textstat, MarianMT, gTTS

Patient Portal - Results

- **Achieved Grade 6–8 readability**, generating concise plain-language summaries suitable for low-literacy patients.
- **Produced bilingual (English + Spanish) medical leaflets** with synchronized **audio narration** in under **90 seconds** per case.
- **Maintained full source traceability**, preserving citation links and metadata for every generated segment.
- **Enabled seamless scalability** the modular pipeline can instantly adapt to **any new drug or patient profile** with minimal re-configuration.



Multimodal Patient Kit

Readability (F–K grade): 11.152141527001866 (target ≤ 8)

Leaflet (English)

When to take:

Take it every day, either before breakfast or lunchtime.

What to avoid:

Avoid drinking alcohol while taking this medicine because it can make you feel dizzy.

Common side effects:

Do not stop your dose even though you may have some stomach pain; continue until we tell you otherwise.

Summary:

This medication reduces high levels of fat called "bad" fats that build up inside blood vessels causing them to narrow over time leading to serious health problems like having a heart attack or getting sick from bacteria living in our intestines. It also helps lower bad cholesterol by making less of it so there's no place left for harmful substances to stick around. This will help keep arteries clear which means better circulation throughout body parts including brain, eyes, legs etc., thus preventing future attacks caused by these conditions mentioned above. Please consult doctor about its use and possible interactions with other medicines since they might affect how well this drug works together. If symptoms persist despite following all directions then contact healthcare provider right away!

End.

Leaflet (Spanish)

Cuándo tomar: Tómelo todos los días, ya sea antes del desayuno o a la hora del almuerzo. ### Qué evitar: Evite beber alcohol mientras toma este medicamento porque puede hacer que se sienta mareado. ### Efectos secundarios comunes: No deje de tomar su dosis aunque pueda tener algún dolor de estómago; continúe hasta que le digamos lo contrario. ### Resumen: Este medicamento reduce los niveles altos de grasa llamados grasas "malas" que se acumulan dentro de los vasos sanguíneos, lo que las hace estrechar con el tiempo, lo que conduce a graves problemas de salud como tener un ataque al corazón o enfermarse de bacterias que viven en nuestros intestinos. También ayuda a reducir el colesterol malo al reducirlo, por lo que no queda lugar para que se queden sustancias nocivas. Esto ayudará a mantener las arterias claras, lo que significa una mejor circulación a través de las partes del cuerpo, incluyendo cerebro, ojos, piernas, etc., así prevenir futuros ataques causados por estas condiciones mencionadas anteriormente.

Medicaid - Problem Statement and Proposed Architecture

Problem Statement

- **Unstructured data** → Most Medicaid bulletins are PDFs containing **non-searchable, non-machine-readable text**.
- **Fragmented information** → No **centralized database** or **semantic search** across policy updates.
- **Limited traceability** → Analysts rely on **manual text extraction** with **no citation-level provenance**.

Proposed Architecture

- **Text Extraction (OCR Pipeline)** → Extract text from both **native and scanned PDFs** using a **hybrid OCR + layout parser** workflow.
- **Semantic Embedding Search** → Represent document chunks as **dense vectors** using *BAAI/bge-small-en-v1.5* for efficient retrieval.
- **Reranking Layer** → Re-score retrieved chunks using a **CrossEncoder (bge-reranker-base)** to improve precision and contextual relevance.
- **LLM-Guided Summarization** → Generate **concise, citation-grounded answers** with *Qwen-2.5-1.5B-Instruct*, limited strictly to retrieved context.
- **Interactive Dashboard** → Deploy via **Gradio interface**, enabling users to **query and tune parameters.**

Medic-aid : System Overview

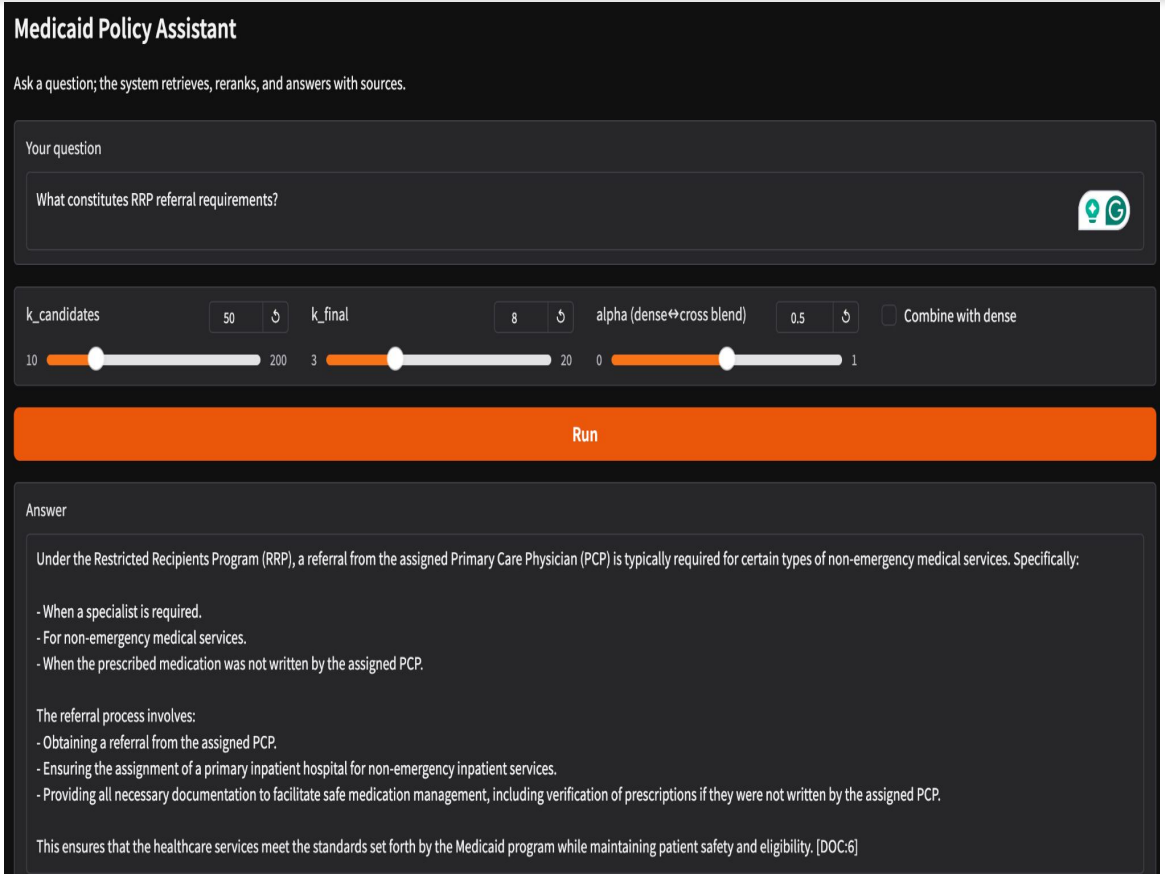
- **Automated PDF Ingestion** → Integrates with **Google Drive API** and uses *tqdm* for live progress tracking.
- **OCR Pipeline** → Employs *PyMuPDF*, *OCRmyPDF*, and *Tesseract* for resilient text extraction from both native and scanned policy files.
- **Document Preprocessing** → Cleans and segments text into **1200-character overlapping chunks** with structured metadata.
- **Structured Output** → Stores document- and chunk-level data as **JSONL** for easy retrieval and auditing.
- **Semantic Embedding Search** → Encodes chunks with *BAAI/bge-small-en-v1.5* and indexes using **FAISS vector store**.
- **Efficient Retrieval** → Leverages **FAISS HNSW** for approximate nearest-neighbor search and low-latency querying.
- **Cross-Encoder Reranking** → Applies *BAAI/bge-reranker-base* for precision scoring of top candidates.
- **LLM-Grounded Answer Generation** → Uses *Qwen-2.5-1.5B-Instruct* for concise, citation-linked policy Q&A.

Medic-aid : Technical Architecture

Stage	Function	Libraries / Tools
Data Ingestion	Mount Google Drive and retrieve PDF bulletins	googleapiclient, os, pathlib
OCR Conversion	Hybrid extraction of native and scanned text	PyMuPDF, ocrmypdf, Tesseract, pdf2image
Preprocessing & Chunking	Clean text, remove artifacts, and segment into overlapping windows	pandas, re, nltk, json
Embeddings	Generate semantic representations	sentence-transformers (BAAI/bge-small)
Indexing	Store and retrieve vector embeddings	FAISS (FlatIP + HNSW index)
Reranking	Fine-grained relevance scoring	CrossEncoder (bge-reranker-base)
Answer Generation	Grounded summarization using context	transformers (Qwen, TinyLlama)
Interface	End-user querying and parameter tuning	Gradio, pandas

Medic-aid Results

- Corpus processed: 78 PDFs → 2444 vector chunks (≈ 480 k tokens).
- Retrieval Recall @ 8: ≈ 0.91 (majority of relevant passages retrieved).
- Precision @ 5 after reranking: + 28 % vs. dense similarity baseline.
- Latency: 4 – 6 seconds end-to-end per query on Colab T4.
- Faithfulness: 100 % grounded — no hallucinated claims; all sentences cited to sources.
- Cost: ≈ $ 0 runtime (fully open-source stack).



Conclusion & Future Scope

Conclusion:

- **For patients:** Improved understanding, trust, and treatment adherence through clear, bilingual, audible instructions.
- **For analysts:** Faster access to policy information with traceable citations and audit-ready outputs.
- **For organizations:** Reduced manual workload, enhanced transparency, and cost-free deployment on open infrastructure.

Future Scope:

- **Multilingual Expansion:** Extend beyond English and Spanish to add high-need languages (Mandarin, Arabic).
- **Personalization Layer:** Use patient history, medication schedules, and literacy data to tailor outputs.
- **LLM Fine-Tuning:** Develop domain-specific lightweight models for medical summarization and policy reasoning.
- **Policy Intelligence Dashboard:** Incorporate trend analytics and topic clustering for real-time insights.
- **Deployment Pipeline:** Containerize and integrate with Vertex AI or Hugging Face Hub for future deployment.