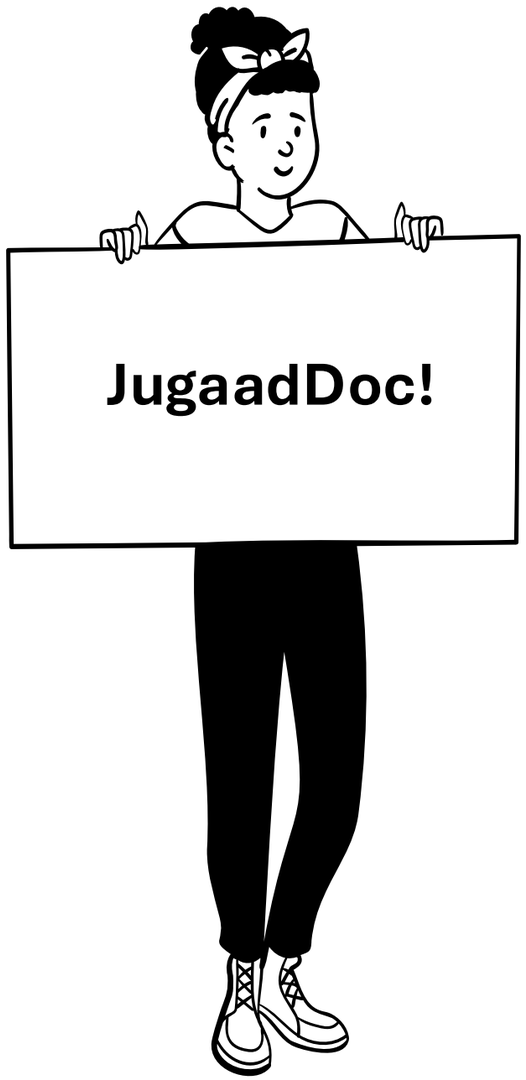
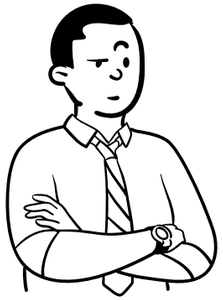
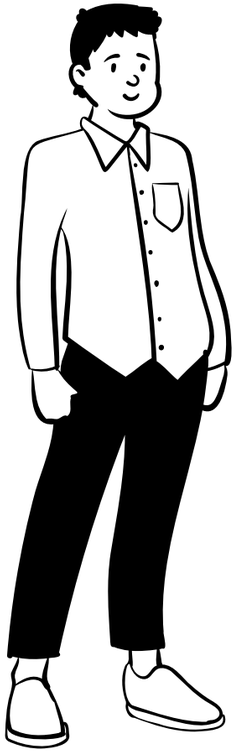


Sindhuja Amudalapadu



Hemanth Kona

Vamshi Krishna Kanisetty



Avinesh Agrawal

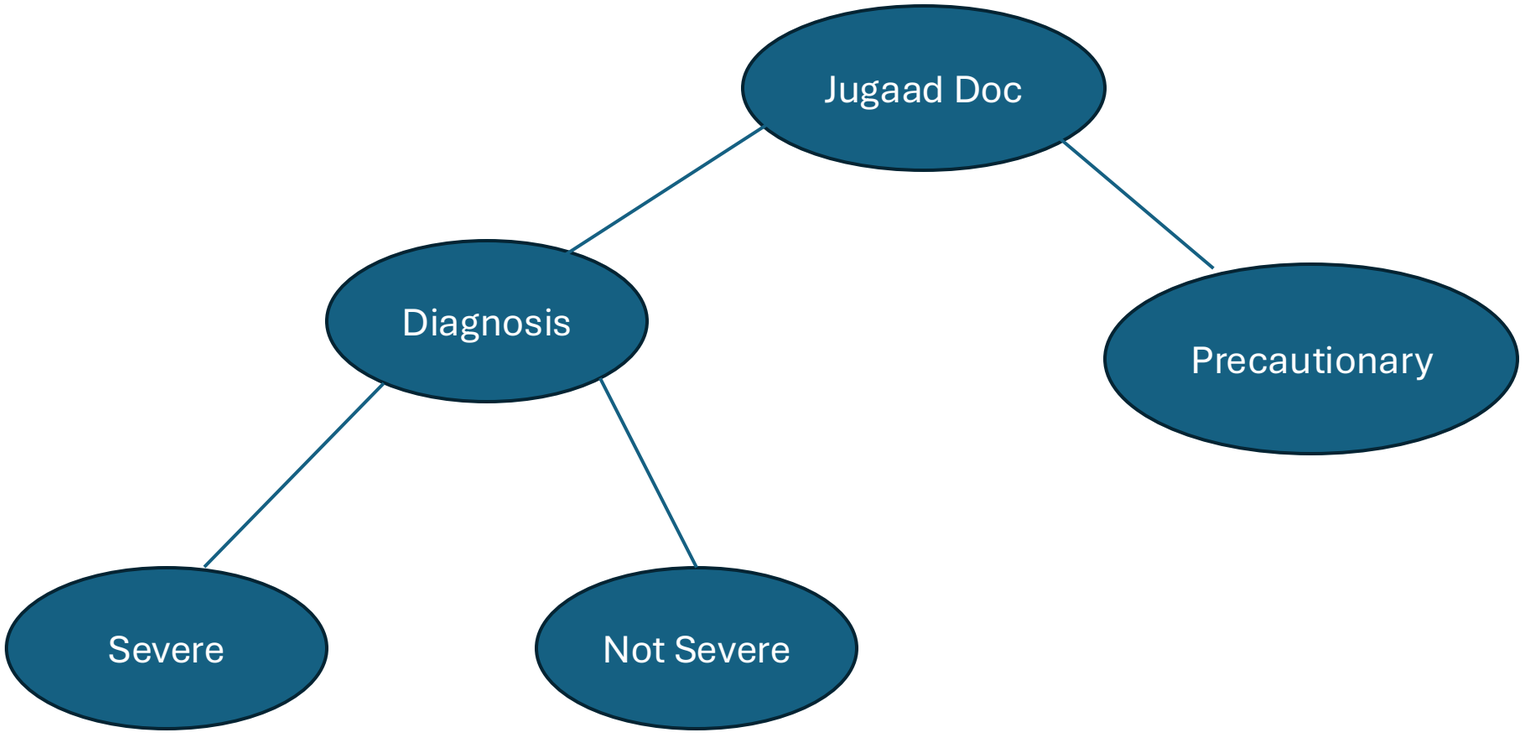
Business Problem

•**Healthcare's digital transformation** is essential to improve patient engagement, response time, and reduce operational costs.

•Traditional methods of interaction are slow and inefficient, leading to the need for an **automated system** to respond to patient inquiries quickly and accurately.

Challenges:

- Limited access to qualified medical care.
- Geographic barriers to healthcare facilities.
- High costs of traveling to urban centers for treatment.



Data Sourcing

Data Collection

Data is collected from various channels:

1. Gale Encyclopedia of Medicine .
2. Gale Encyclopedia of Alternate Medicine

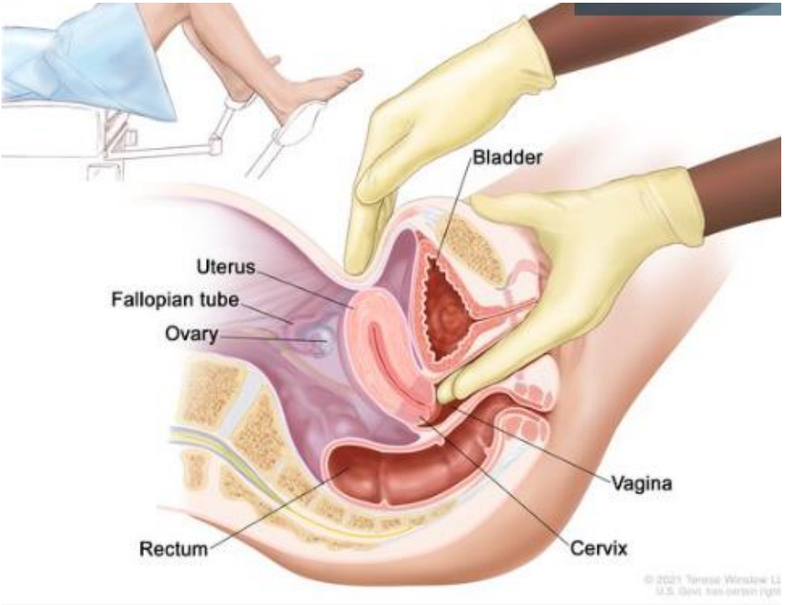
Key Data Includes

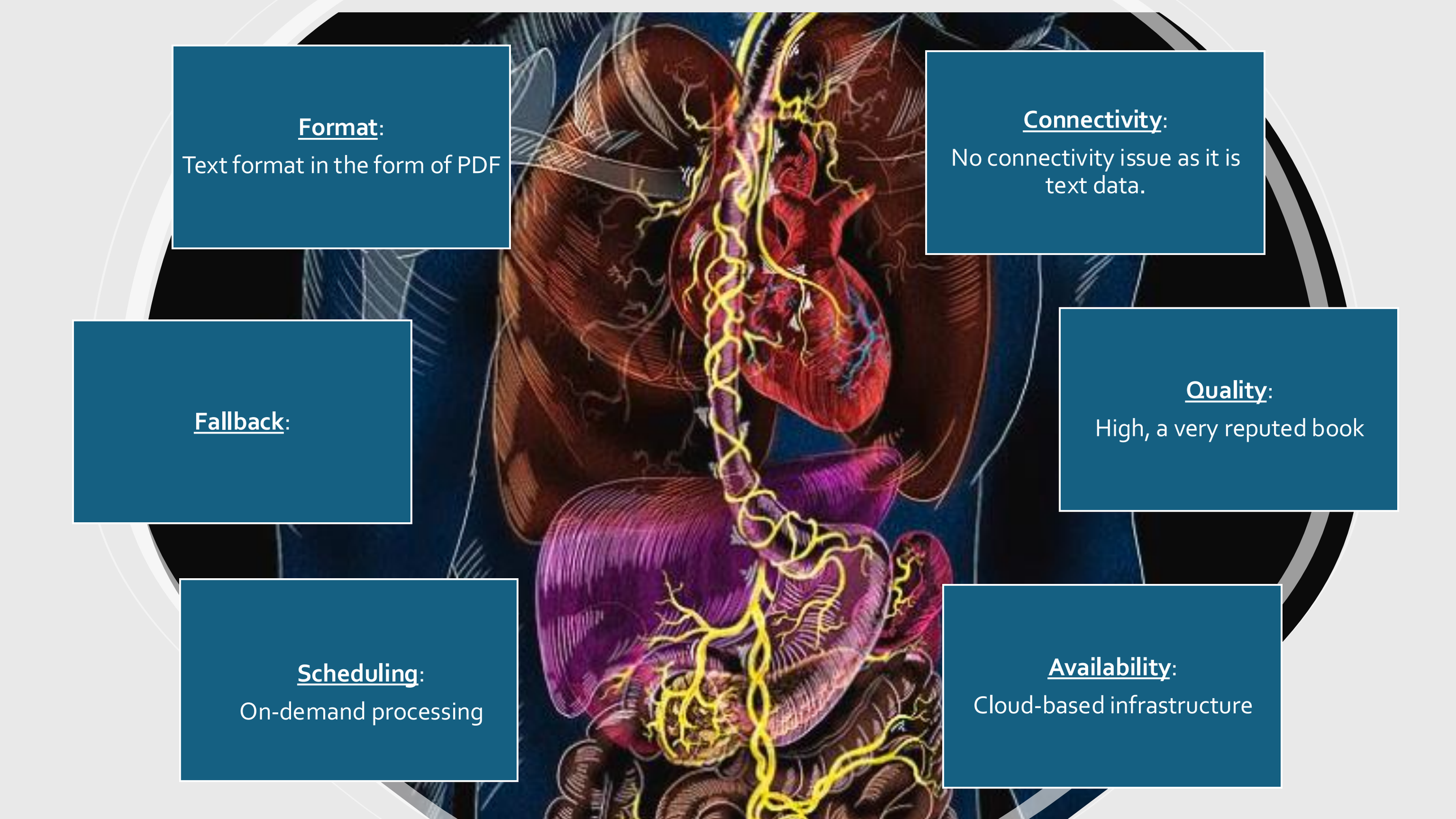
The key data encompasses information related to various medical specialties, including gynecology, cardiology, neurology, pulmonology, ophthalmology, and other departments, along with their associated problems and solutions.

SCOPE

Almost 1,700 full-length articles are included in the *Gale Encyclopedia of Medicine 2*, including disorders/conditions, tests/procedures, and treatments/therapies. Many common drugs are also covered, with generic drug names appearing first and brand names following in parentheses, eg. acetaminophen (Tylenol). Throughout the *Gale Encyclopedia of Medicine 2*, many prominent individuals are highlighted as sidebar biographies that accompany the main topical essays. Articles follow a standardized format that provides information at a glance. Rubrics include:

Disorders/Conditions	Tests/Treatments
Definition	Definition
Description	Purpose
Causes and symptoms	Precautions
Diagnosis	Description
Treatment	Preparation
Alternative treatment	Aftercare
Prognosis	Risks
Prevention	Normal/Abnormal results
Resources	Resources
Key terms	Key terms





Format:
Text format in the form of PDF

Connectivity:
No connectivity issue as it is text data.

Fallback:

Quality:
High, a very reputed book

Scheduling:
On-demand processing

Availability:
Cloud-based infrastructure

Data Quality

Completeness

Includes necessary medical data

Integrity

Content remains intact and unaltered. Digital signatures or encryption techniques are often used .

Confirmity

Includes standardized units and ranges to avoid misinterpretation.

Accuracy

For medical data, accuracy is crucial as incorrect information can lead to wrong diagnoses, treatments, or prescriptions.

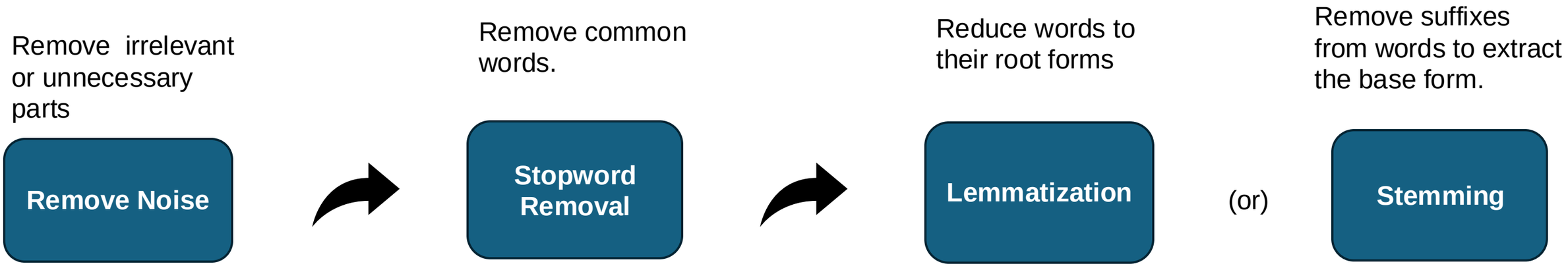
Consistency

same medical terms, patient identifiers, treatment protocols, and diagnostic information should be represented uniformly

Timeliness

Documents reflects the most current diagnoses, treatments

Data Cleaning



ETL Process and Data Storage

1

Extract

Data is collected from various sources such as PDFs, APIs, and online repositories.

2

Transform

Preprocessing tasks like lemmatization, stemming, and stopword removal are performed.

3

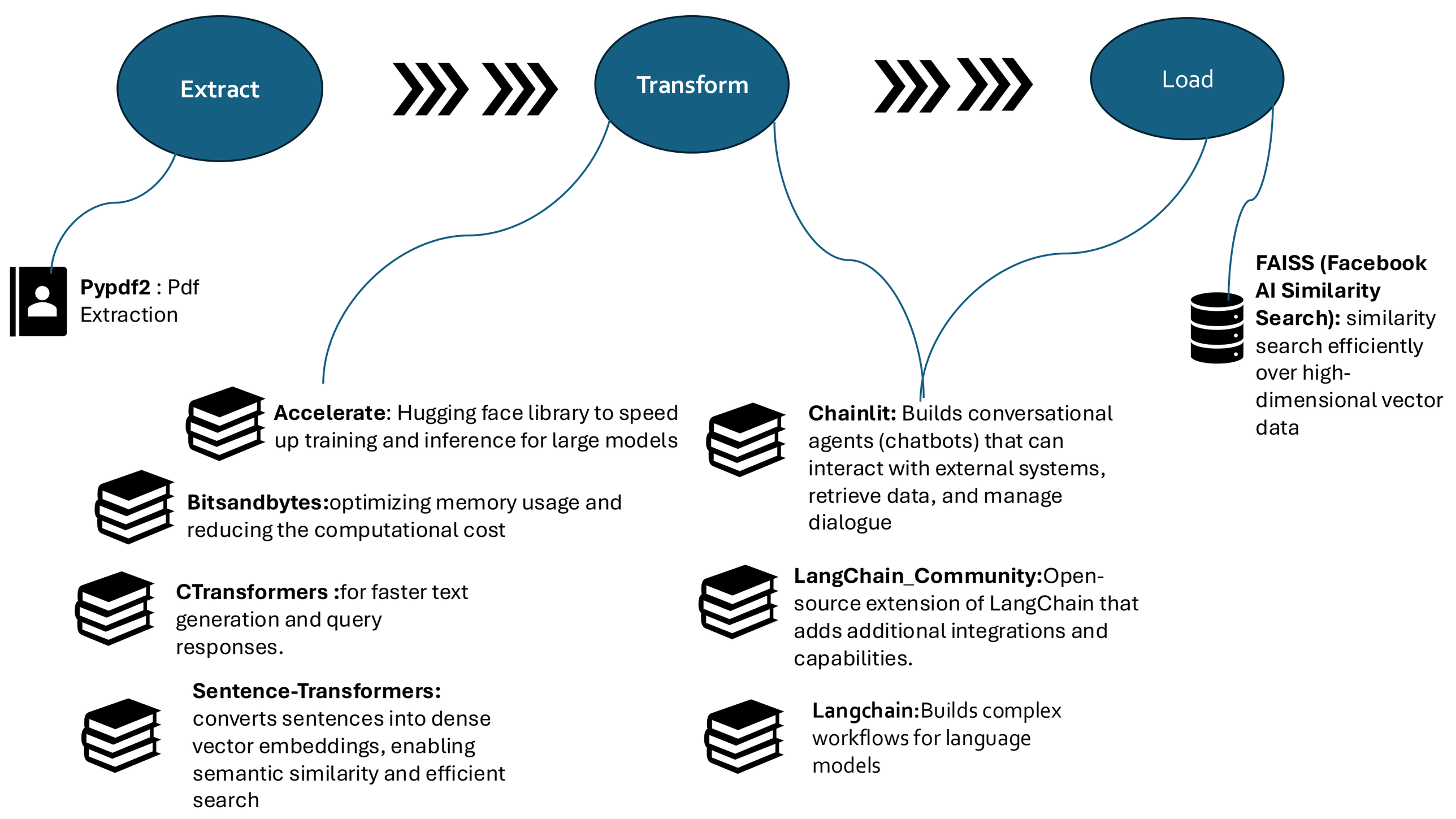
Load

Transformed data is loaded into AWS cloud storage solutions like S3, RDS, and DynamoDB.

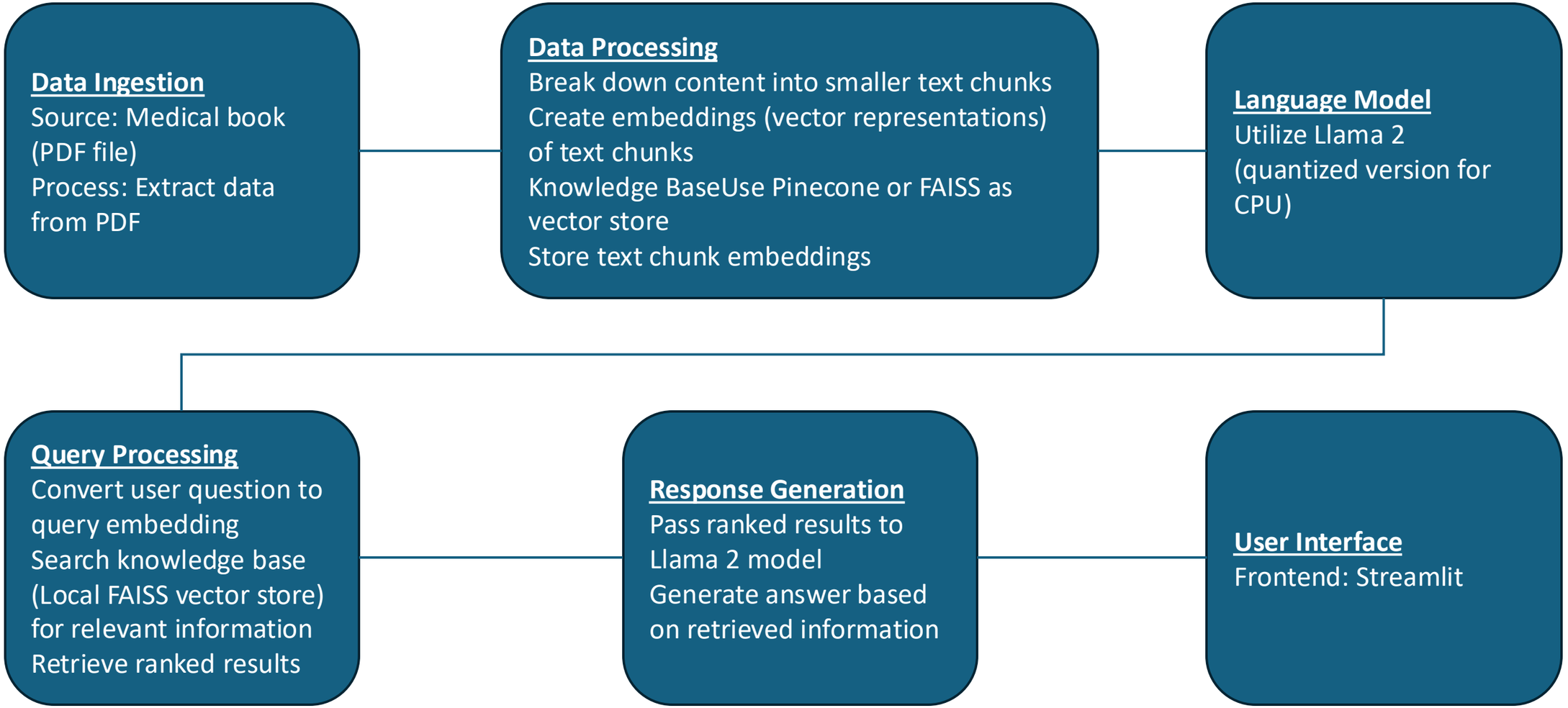
4

Backup and Security

Backup and security measures ensure data durability, availability, and compliance with regulations.


Extract
Transform
Load
Pypdf2 : Pdf Extraction
Accelerate: Hugging face library to speed up training and inference for large models
Bitsandbytes:optimizing memory usage and reducing the computational cost
CTransformers :for faster text generation and query responses.
Sentence-Transformers: converts sentences into dense vector embeddings, enabling semantic similarity and efficient search
Chainlit: Builds conversational agents (chatbots) that can interact with external systems, retrieve data, and manage dialogue
LangChain_Community:Open-source extension of LangChain that adds additional integrations and capabilities.
Langchain:Builds complex workflows for language models
FAISS (Facebook AI Similarity Search): similarity search efficiently over high-dimensional vector data

Application Flow



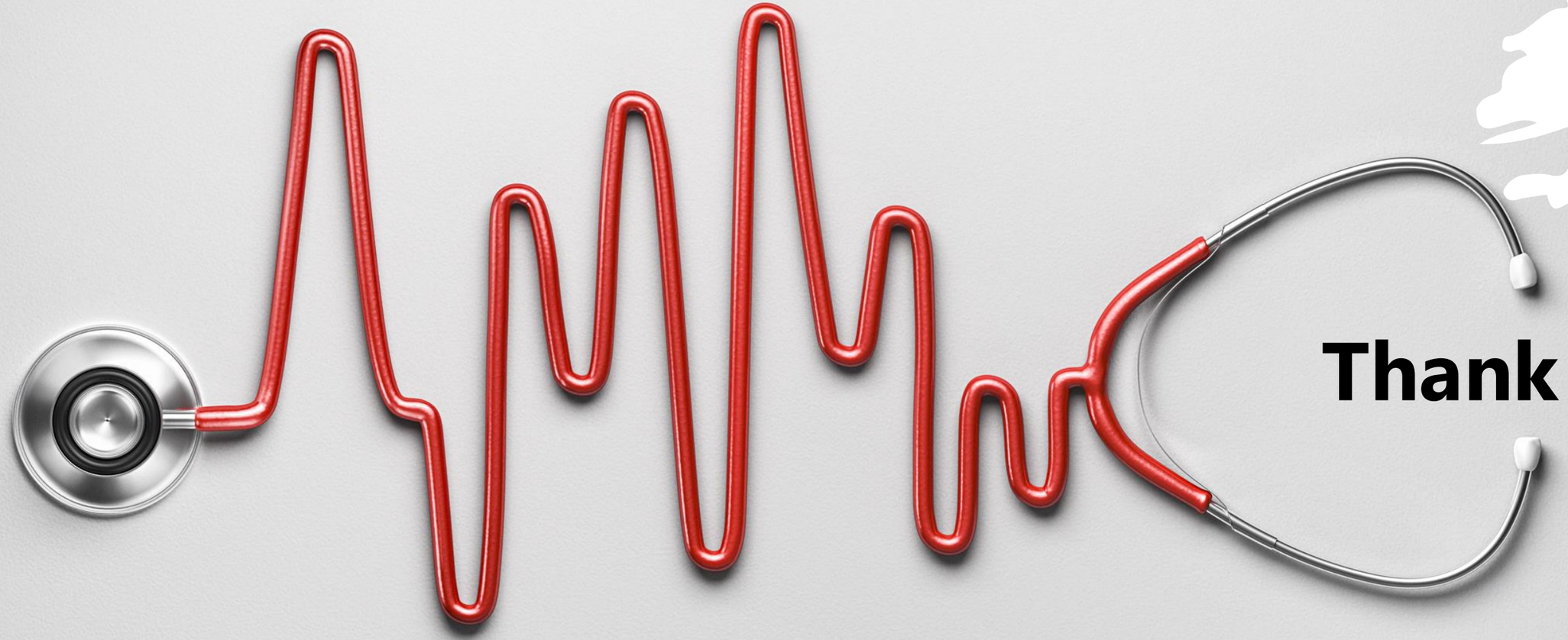
Conclusion & Next Steps

The project aims to build a robust medical chatbot using AI-driven text processing, ETL pipelines, and advanced retrieval-augmented generation.

Prototype testing with medical professionals will validate the chatbot's capabilities.

Scaling the chatbot's functionality for multiple use cases is a key next step.

Ongoing improvements based on user feedback and data quality monitoring will ensure continuous enhancement.



Thank You