# TrustMed AI: An AI-Powered Chatbot for the Healthcare Industry

Vishnu Menon
*Arizona State University*
Tempe, AZ 85281
vmenon@asu.edu

Shitij Mathur
*Arizona State University*
Tempe, AZ 85281
skmathu1@asu.edu

Advaith Venkatsubramanian
*Arizona State University*
Tempe, AZ 85281
avenk138@asu.edu

Suhas Gajula
*Arizona State University*
Tempe, AZ 85281
sgajul11@asu.edu

Thanishka Bolisetty
*Arizona State University*
Tempe, AZ 85281
tboliset@asu.edu

Varad Vitthal More
*Arizona State University*
Tempe, AZ 85281
vmore4@asu.edu

***Abstract*—The exponential growth of online healthcare discussions will make it increasingly challenging for patients to distinguish between reliable and misleading information. *TrustMed AI* will be developed as an intelligent healthcare chatbot designed to address this issue by integrating retrieval-augmented generation (RAG), medical ontologies, and explainability frameworks to ensure accuracy, transparency, and trust. The system will be implemented using LangChain and evaluated with TruLens, combining structured knowledge bases such as UMLS and MetaMap with real-world patient interactions from verified medical communities. *TrustMed AI* will be capable of retrieving, reasoning, and generating grounded medical responses that are contextually relevant and evidence-based, ultimately improving user confidence and access to verified healthcare information.**

***Index Terms*—LangChain, TruLens, Healthcare Chatbot, Retrieval-Augmented Generation, UMLS, Ontology Integration, Explainable AI**

## I. Problem Definition

The increasing dependence on digital platforms for medical guidance will make it difficult for individuals to identify accurate and trustworthy health information. Online sources such as discussion forums, blogs, and social media will continue to include unverified claims, anecdotal experiences, and conflicting advice that can mislead patients. This growing challenge highlights the need for an intelligent system capable of delivering medically grounded, contextually appropriate, and transparent responses to user queries.

*TrustMed AI* will aim to bridge this gap by developing a healthcare chatbot that can retrieve and reason over verified medical information while maintaining interpretability and factual accuracy. The system will combine natural language understanding with structured medical knowledge to provide reliable, evidence-based responses that enhance patient trust and accessibility. By doing so, it will contribute to safer and more informed online healthcare interactions.

## II. Data Sets

To build a comprehensive and reliable knowledge base for the TrustMed AI chatbot, we gather data from three distinct, complementary categories of sources. This multi-faceted approach ensures that the AI's understanding is grounded in clinical evidence, aligned with real-world patient queries, and structured for semantic consistency.

### *A. Authoritative Medical Sources*

The foundation of our dataset consists of content from highly reputable, evidence-based medical sources. This includes peer-reviewed publications from top-tier journals such as the New England Journal of Medicine (NEJM) and the Journal of the American Medical Association (JAMA), as well as physician-vetted articles and clinical guidelines from trusted organizations like the Mayo Clinic and WebMD. This content provides the core of clinically accurate and up-to-date medical facts that underpin the chatbot's responses.

### *B. Online Health Forums and Communities*

To capture the language and context of real-world patient concerns, we incorporate data from public online health communities. Sources include specific Reddit subreddits (e.g., r/AskDocs, r/Diabetes), patient support networks like PatientsLikeMe and SmartPatients, and moderated forums such as Mayo Clinic Connect. This user-generated content is invaluable for training the AI to understand colloquial terminology, common questions, and the practical, lived experiences of patients. All community data undergoes a rigorous curation process to filter for relevance and quality.

### *C. Medical Ontologies and Semantic Resources*

To unify the diverse terminology across our datasets, we leverage structured medical knowledge bases. The primary resource is the Unified Medical Language System (UMLS) from the U.S. National Library of Medicine [3]. By using tools like MetaMap [4] and the SPECIALIST NLP Lexicon, we perform entity linking to map textual mentions (e.g., "heart attack," "high blood sugar") to standardized UMLS concepts (e.g., Myocardial Infarction, Hyperglycemia). This creates a consistent, semantically linked knowledge graph that enables the chatbot to understand medical concepts regardless of how they are phrased.

All data is collected using automated web scraping tools (e.g., Selenium, Playwright) and subjected to a thorough cleaning and preprocessing pipeline. The critical step in this process is standardizing all information by linking identified medical entities to their corresponding UMLS concepts. The final output is a unified, high-quality dataset that serves as the knowledge repository for the TrustMed AI chatbot.

## III. State-of-the-Art Methods and Algorithms

Recent advancements in conversational AI and biomedical information retrieval have paved the way for intelligent systems that can synthesize and deliver evidence-based medical information in natural language. Large Language Models (LLMs) such as OpenAI's GPT-4 and Google's Med-PaLM have demonstrated exceptional capability in understanding and generating health-related text, though their reliability depends on curated, authoritative data sources. Current healthcare chatbots (e.g., WebMD's Symptom Checker, Ada Health, Babylon) primarily provide static, symptom-based guidance or prescripted decision trees, which limits their adaptability and contextual awareness.

State-of-the-art research emphasizes combining retrieval-augmented generation (RAG) and knowledge-grounded conversational systems to improve trustworthiness and transparency in medical dialogue. Efforts such as LangChain and LangGraph frameworks enable modular workflows that integrate data ingestion, retrieval, reasoning, and response generation. TrustMed AI aims to advance this paradigm by continuously ingesting content from authoritative sources (e.g., NEJM, JAMA, Mayo Clinic, WebMD), structuring it into a unified medical knowledge base, and linking it to patient community discussions. The project differentiates itself by grounding conversational responses in verifiable citations, enhancing the interpretability and reliability of health advice while maintaining accessibility through natural language interaction.

The core architecture of TrustMed AI leverages a pipeline combining web scraping, semantic structuring, retrieval-augmented reasoning, and evaluation. Data acquisition will be handled via Selenium, BeautifulSoup or Playwright, enabling automated extraction of structured and unstructured data from trusted medical sources and healthcare forums. Extracted content will be preprocessed through Named Entity Recognition (NER). The agentic workflow will be orchestrated using LangChain or LangGraph, enabling modular components for data retrieval, context management, and reasoning. Specifically, a RAG architecture will be employed: relevant information is retrieved from the structured medical knowledge base using dense embeddings (e.g., OpenAI or HuggingFace biomedical models such as BioBERT[2] or PubMedBERT), then passed to an LLM that synthesizes coherent, context-aware responses with explicit citations. Conversation management modules will handle dialogue context retention and multi-turn query understanding.

Evaluation will follow both quantitative and qualitative frameworks. TruLens will be used to measure factual accuracy, faithfulness, and citation alignment of responses. Precision and recall metrics will assess the retrieval effectiveness of relevant medical facts, while human evaluators will review interpretability and user trust. Together, these algorithmic and analytical components ensure that TrustMed AI remains both clinically reliable and conversationally natural.

## IV. Research and Development Plan

The research and development (R&D) of TrustMed AI will follow a structured, phased approach to ensure measurable progress, systematic evaluation, and iterative refinement as shown in Figure 1. The plan integrates techniques from biomedical NLP, ontology-driven retrieval, and conversational AI engineering, ensuring both academic rigor and practical feasibility.

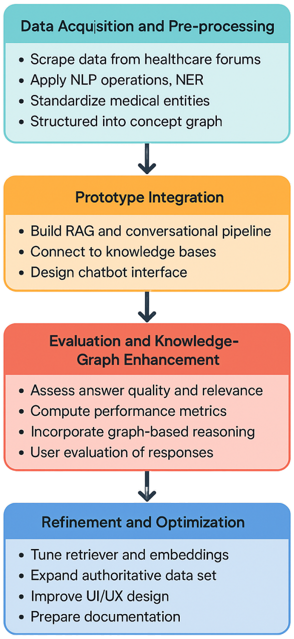
Data Acqu|sition and Pre-processing
• Scrape data from healthcare forums
• Apply NLP operations, NER
• Standardize medical entities
• Structured into concept graph
Prototype Integration
• Build RAG and conversational pipeline
• Connect to knowledge bases
• Design chatbot interface
Evaluation and Knowledge-Graph Enhancement
• Assess answer quality and relevance
• Compute performance metrics
• Incorporate graph-based reasoning
• User evaluation of responses
Refinement and Optimization
• Tune retriever and embeddings
• Expand authoritative data set
• Improve UI/UX design
• Prepare documentation

Fig. 1: R&D Workflow

### A. Data Acquisition and Pre-processing

The initial stage of the project is dedicated to building the foundational dataset and ontology mappings that will support subsequent retrieval and reasoning tasks. Data will be scraped from verified healthcare forums such as r/AskDocs, PatientsLikeMe, HealthBoards, and Mayo Clinic Connect using Selenium and BeautifulSoup4 to handle dynamic content extraction. Once collected, the raw corpus will undergo a series of natural language processing (NLP) operations, including tokenization, stop-word removal, and lemmatization, in order to normalize and clean the text for downstream analysis. To extract domain-specific medical entities, NER will be applied using biomedical models such as SciSpacy and BioBERT. These extracted entities will then be standardized against authoritative resources, specifically the UMLS Metathesaurus and MetaMap, to map free-text mentions to Concept Unique Identifiers (CUIs). Finally, the processed data will be structured into a concept graph that encodes the relationships among diseases, symptoms, and treatments. By the end of this phase, the system will have produced a clean, structured dataset together with ontology-based entity mappings, forming a robust knowledge layer for accurate and clinically aligned medical information retrieval.

### *B. Prototype Integration*

The second phase of development focuses on implementing the retrieval and generation pipeline that forms the basis of the functional prototype. Textual data is first encoded using domain-specific biomedical embeddings such as PubMedBERT, BioBERT, or BioClinicalBERT, which are stored in a PGVector database to enable efficient semantic search. Retrieval operations employ Approximate Nearest Neighbor (ANN) algorithms, to achieve scalability while maintaining high recall and precision. On top of this layer, a RAG pipeline is constructed, where user queries are processed through a dense retriever, passed through a context selector, and then synthesized into coherent responses by a large language model such as Llama 3.1 or Qwen 2.5. This architecture is orchestrated using LangChain or LangGraph to ensure modularity and experimental flexibility, while incorporating context window management and multi-turn dialogue tracking for realistic conversational flow. A preliminary user interface is then developed using Streamlit or a lightweight Flask/React.js frontend, equipped with LangChain memory components to handle session-based context retention. The outcome of this phase is a prototype chatbot that can retrieve authoritative information, generate responses grounded in citations, and demonstrate the feasibility of the pipeline.

### *C. Evaluation and Knowledge-Graph Enhancement*

The third phase introduces a rigorous evaluation framework and knowledge graph enhancements. Quantitative assessment is conducted using automated metrics including precision, recall, F1-score for retrieval accuracy, and query latency for performance benchmarking. Explainability metrics are captured using TruLens, which provides groundedness, answer relevance, and context relevance scores. Complementing these, human-in-the-loop evaluation is performed to measure interpretability and user trust. To strengthen reasoning and reduce hallucinations, the baseline RAG pipeline is extended with knowledge graph augmentation through GraphRAG. This involves extracting subject–predicate–object triples from text using LLM-powered extraction methods (e.g., GraphRAG or LangExtract, as described in the KG Extraction tutorial), and constructing a biomedical knowledge graph in Neo4j for structured query augmentation. Advanced graph-based inference is further supported by Graph Convolutional Networks (GCNs) to capture latent entity relationships. Iterative ablation studies are conducted to compare vanilla RAG, GraphRAG, and ontology-only retrieval approaches, with retriever thresholds (e.g., top-k retrieval) tuned to optimize the trade-off between groundedness and system latency. This phase produces an evaluation report with detailed metrics and comparative insights into the impact of graph-based reasoning on chatbot reliability.

### *D. Refinement and Optimization*

The final phase centers on refinement, optimization, and preparation of deliverables. System performance is optimized through retriever tuning, which involves adjusting similarity thresholds, embedding selection, and ANN parameters. Prompt engineering techniques are introduced, employing structured templates that enforce citation requirements, ontology alignment, and reasoning chains. To reduce inference cost and improve throughput, caching mechanisms such as vector cache layers are implemented. The dataset is expanded by ingesting additional authoritative medical sources, including NEJM, JAMA, and WebMD APIs, and continual embeddings updates are applied to support incremental learning. User interface enhancements focus on trust and interpretability, such as visualizing groundedness scores from TruLens in a color-coded format and embedding inline citations linked directly to source ontology identifiers or URLs. The final deliverables from this phase include an optimized, explainable chatbot system, comprehensive technical documentation, and a demonstration-ready prototype. Collectively, these refinements ensure that TrustMed AI is robust, scalable, and deployable in real-world healthcare information retrieval settings.

## V. Evaluation Plan

Our evaluation strategy is centered around rigorously assessing the technical performance of the TrustMed AI system, particularly through the integration of TruLens for in-depth analysis. This plan is structured to ensure that the system effectively meets its goal of providing accurate, grounded, and contextually relevant healthcare information retrieval.

### *A. Technical Performance Metrics*

We utilize a set of quantitative and qualitative metrics to evaluate the system's performance, focusing on:

- **TruLens Analysis:** At the center of our evaluation framework, TruLens serves as the principal tool for assessing *Groundedness*, *Answer Relevance*, and *Context Relevance* [1]. This analytical approach enables a systematic examination of the extent to which the system's outputs are anchored in factual information extracted from the underlying database and their alignment with user queries. TruLens is seamlessly integrated within the LangChain agent to capture comprehensive traces of model behavior, including inputs, tool interactions, intermediate reasoning processes, and generated outputs. The feedback functions embedded in TruLens yield interpretable metrics that elucidate the model's relative strengths and limitations in producing factually consistent and contextually appropriate healthcare information. Figure 2 depicts the feedback mechanisms employed in this study to facilitate a qualitative evaluation of the retrieval agent's performance.

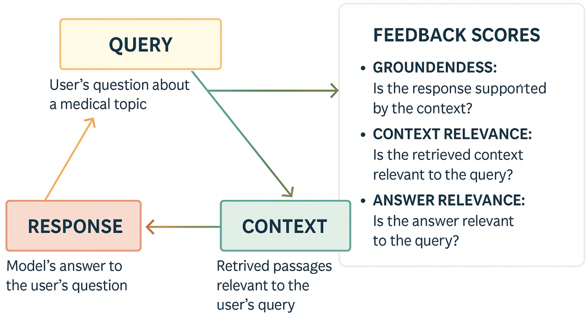
QUERY
User's question about a medical topic
FEEDBACK SCORES
• GROUNDENDESS:
Is the response supported by the context?
• CONTEXT RELEVANCE:
Is the retrieved context relevant to the query?
• ANSWER RELEVANCE:
Is the answer relevant to the query?
RESPONSE
Model's answer to the user's question
CONTEXT
Retrived passages relevant to the user's query

Fig. 2: TruLens Analysis Process Cycle showing interconnected stages of Groundedness, Answer Relevance, and Context Relevance.

- **Precision and Recall Analysis:** To quantitatively assess the accuracy of information retrieval, precision and recall metrics are employed. Precision measures the proportion of correctly retrieved medical facts among all results returned by the system, while recall evaluates how many relevant facts were successfully identified. These metrics are computed as:

$$\text{Precision} = \frac{TP}{TP + FP}, \qquad \text{Recall} = \frac{TP}{TP + FN} \tag{1}$$

  where $TP$ denotes true positives, $FP$ false positives, and $FN$ false negatives. Precision reflects the reliability of the system's retrieved information, while recall captures its completeness. The harmonic mean (F1-score) of these two values serves as a balanced measure of overall performance, ensuring that TrustMed AI maintains both accuracy and coverage in healthcare information retrieval.

### B. Iterative Refinement Process

Feedback derived from the technical performance evaluation will inform the ongoing refinement of our system. This iterative process includes:

- Fine-tuning and enhancing the algorithmic framework using insights from TruLens metrics to boost the precision and contextual relevance of retrieved information.
- Regularly updating the data collection and retrieval components to integrate the most recent healthcare information and user feedback, thereby ensuring continuous improvements in system performance, reliability, and trustworthiness.

## VI. Project Timeline: Tasks, Descriptions, and Deadlines

To ensure systematic progress and timely completion, the project is divided into distinct phases with clearly defined tasks, durations, and deliverables as shown in Table I. The timeline below outlines the major milestones, expected outcomes, and deadlines associated with each phase of development.

TABLE I: Project Timeline and Deliverables

| Phase | Weeks | Major Tasks | Deliverables |
|---|---|---|---|
| 1 | 1–3 | Data scraping and ontology preprocessing | Clean dataset and entity mapping |
| 2 | 4–6 | Develop RAG pipeline and prototype UI | Working chatbot prototype |
| 3 | 7–8 | Integrate TruLens and evaluate responses | Evaluation metrics and analysis report |
| 4 | 9–10 | System refinement and documentation | Final presentation and report submission |

## VII. Division of Work

The project tasks were distributed among the team members to align with their expertise. Table II outlines the key responsibilities assigned to each member.

TABLE II: Team Roles and Responsibilities

| Member | Core Responsibilities |
|---|---|
| Vishnu Menon | System architecture, RAG pipeline orchestration, and prompt engineering |
| Varad Vitthal More | Data scraping and development of the data ingestion pipeline |
| Advaith Venkatsubramanian | Ontology integration, including UMLS linking and data normalization |
| Thanishka Bolisetty | Development of the chatbot user interface (UI/UX) and session management |
| Shitij Mathur | Implementation of the retrieval system and performance optimization |
| Suhas Gajula | System evaluation, including metric analysis (Precision, Recall) and error analysis |

## References

[1] Datta, A., Fredrikson, M., Leino, K., Lu, K., Sen, S., Shih, R., & Wang, Z. (2022, July). Exploring conceptual soundness with trulens. In NeurIPS 2021 Competitions and Demonstrations Track (pp. 302-307). PMLR.

[2] Lee, Jinhyuk & Yoon, Wonjin & Kim, Sungdong & Kim, Donghyeon & Kim, Sunkyu & So, Chan & Kang, Jaewoo. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics (Oxford, England). 36. 10.1093/bioinformatics/btz682.

[3] O. Bodenreider, ”The Unified Medical Language System (UMLS): integrating biomedical terminology,” *Nucleic Acids Research*, vol. 32, no. suppl_1, pp. D267–D270, Jan. 2004.

[4] A. R. Aronson, ”Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program,” in *Proc. AMIA Symp.*, 2001, p. 17.