P1: TrustMed AI
An AI-Powered Chatbot for the Healthcare Industry

A presentation by Group #3:

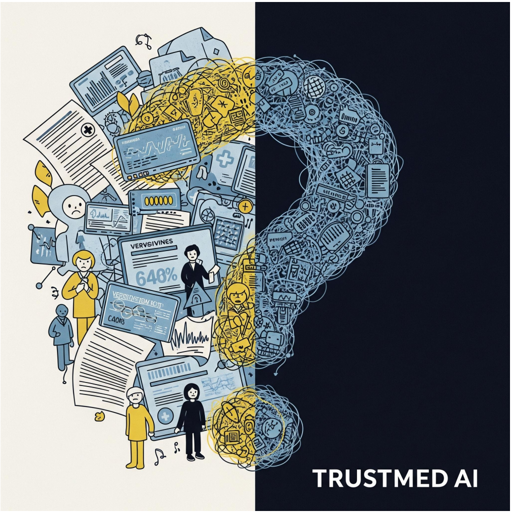
Thanishka Bolisetty | **Shitij Mathur** | **Suhas Gajula**

Varad Vitthal More | **Advaith Venkatsubramanian** | **Vishnu Menon**

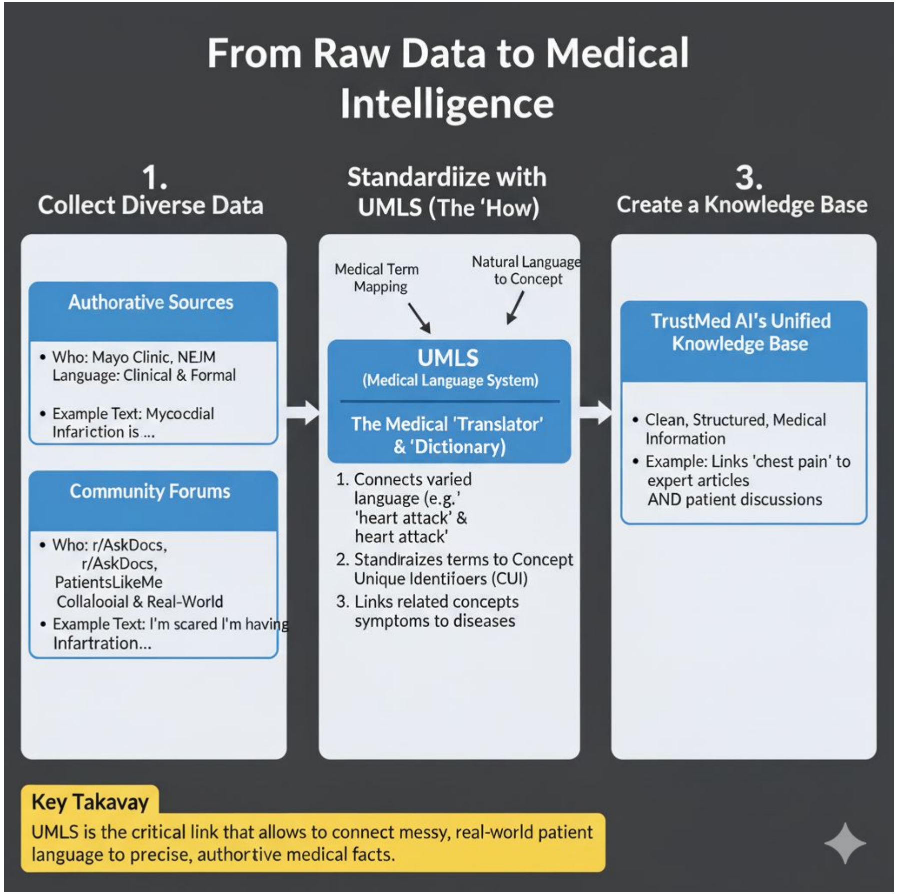
Arizona State University

Problem Definition

- Growing reliance on online health information makes identifying reliable sources difficult.
- Unverified or conflicting advice in online platforms can mislead patients.
- Need for an AI chatbot that understands user queries, retrieves and reasons over verified medical data.
- Goal: Combine a deep understanding of each user with authentic, data-driven insights to deliver healthcare guidance that’s accurate, transparent, and easy to trust.



Dataset Description



Why This Process Matters

- **Solves the "Language Gap":** We bridge the gap between formal medical terminology (like "Myocardial Infarction") and real-world patient language (like "heart attack").
- **Creates a Unified Knowledge Graph:** Instead of having two separate, disconnected datasets, UMLS acts as the "glue." It links a *symptom* from a forum to a *disease* in a medical journal.
- **Enables True Semantic Search:** Our AI doesn't just match keywords. It understands the *medical concept* behind the words. A search for "high blood sugar" will also find expert articles on "Hyperglycemia."
- **Reduces Errors & Hallucinations:** By grounding all data to standardized UMLS concepts, we ensure our AI's responses are based on a consistent, verifiable, and accurate medical framework.
- **The Final Result:** This pipeline transforms raw, messy data into an intelligent, interconnected knowledge base. This is what allows TrustMed AI to retrieve the *right* information and provide a relevant, *grounded* answer.

State-of-the-Art Methods and Algorithms

- Current healthcare chatbots (e.g., WebMD Symptom Checker, Ada Health, Babylon) provide static, symptom-based guidance or prescripted decision trees, which limits their adaptability and contextual awareness.
- Using advances in LLM, we can advance beyond current chatbots by moving from static bots to an intelligent, evidence-based conversational system.
- Leverages a RAG architecture to continuously ingest and structure data from authoritative medical sources into a unified knowledge base.
- Differentiates itself by grounding responses in verifiable citations, dramatically improving interpretability and reliability.
- Utilizes a core technical pipeline of data acquisition, semantic structuring, and retrieval-augmented reasoning orchestrated with LangChain/LangGraph.
- Ensures clinical reliability through rigorous quantitative and qualitative evaluation of factual accuracy and user trust.

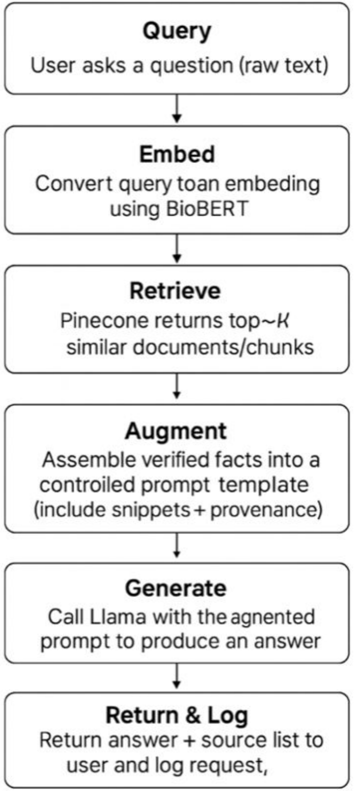
Research and Development Plan

Phase 1: Data Acquisition & Pre-Processing
Scrape verified forums, apply NER, and map to UMLS.

Phase 2: Prototype Integration
Implement RAG pipeline and store embeddings (Pinecone Vector DB).
Develop UI with Streamlit/React.

Phase 3: Evaluation & Graph Enhancement
Use TruLens for evaluation; build Neo4j knowledge graph.

Phase 4: Refinement & Optimization
Tune retrievers, embeddings, and prompts.
Improve UI/UX and integrate citation visualization.



Project Architecture

1. START

- Begins when a user submits a query.

2. JUDGE NODE

- Evaluates the user’s question.
- Determines if external factual data (from the knowledge base) is required.
- **If yes** → routes to the **RAG Node**.
- **If no** → routes directly to the **Answer Node** for a standard LLM response.

3. RAG NODE

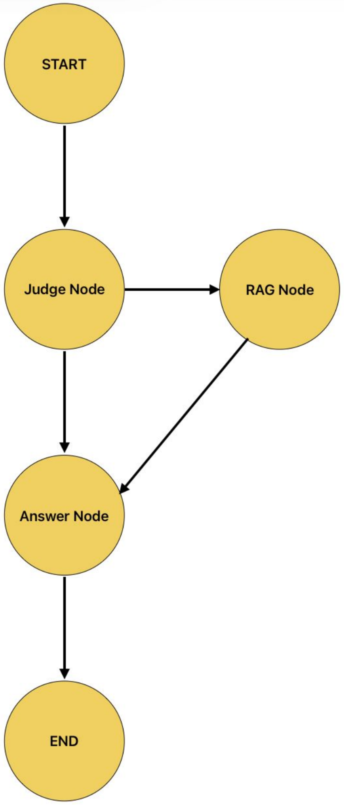
- Performs Retrieval-Augmented Generation.
- Retrieves relevant documents or facts from the **Vector Database (Pinecone)**.
- Integrates retrieved context into the prompt for grounded, fact-based generation.
- Passes the enriched context to the **Answer Node**.

4. ANSWER NODE

- Generates the final response.
- Uses **Llama** as the LLM.
- If RAG data is available → produces a grounded answer.
- If RAG wasn’t used → generates a direct model response.
- Ensures consistent output format and citations (if applicable).

5. END

- Returns the response to the user.
- Logs decision flow, retrieved sources, and latency for monitoring.



Design of Experiments & Evaluation Plan

Goal: Ensure TrustMed AI provides accurate, grounded, and contextually relevant healthcare information.

Evaluation Strategy: Combines **quantitative** (automated metrics) and **qualitative** (human-in-the-loop) assessment.

Core Framework:

- **TruLens Integration** – primary tool for systematic evaluation of model performance.
- Embedded within **LangChain agent** to trace inputs, reasoning steps, and outputs.
- Provides interpretable feedback on **Groundedness**, **Answer Relevance**, and **Context Relevance**.

Data Flow Insight: TruLens feedback cycle identifies strengths and weaknesses of retrieval and generation.

Focus Areas: Faithfulness to medical sources, alignment with user intent, and factual accuracy.

Design of Experiments & Evaluation Plan (ctd)

Quantitative Metrics

- **Precision:** Ratio of correctly retrieved medical facts to total retrieved.
- **Recall:** Ratio of correctly retrieved facts to total relevant facts.
- **F1-Score:** Harmonic mean of Precision and Recall to balance accuracy and coverage.

Qualitative Metrics:

- **Human Evaluation:** Experts review interpretability, reliability, and user trust.
- Visualization of **TruLens groundedness scores** (color-coded) and inline citations for transparency.

Iterative Refinement Process:

- Fine-tune retrieval and reasoning using TruLens feedback.
- Continuously update medical datasets and retriever parameters for improved performance.
- Maintain system **trustworthiness, factuality, and clinical reliability** through periodic re-evaluation.

Deliverables: Evaluation report, performance dashboard, and optimized RAG pipeline.

Project Plan: Tasks, Deadlines & Division of Work

Timeline:

Weeks 1–3: Data scraping & ontology preprocessing → Clean dataset.
Weeks 4–6: RAG pipeline & UI → Working prototype.
Weeks 7–8: TruLens evaluation → Metrics report.
Weeks 9–10: Refinement & documentation → Final presentation.

Team Roles:

Vishnu Menon – Architecture & Prompt Engineering
Varad More – Data Ingestion
Advaith Venkatsubramanian – Ontology Integration
Thanishka Bolisetty – UI/UX
Shitij Mathur – Retrieval Optimization
Suhas Gajula – Evaluation & Metrics

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.

Arizona State University