Team # 3

AI Assisting in Medical Sciences — The MedRAG Framework for Medical Question Answering

Yogesh Yadav, Narahari Kommi

School of Electrical & Computer Engineering, Arizona State University

Contacts: [yyadav8, nkommi]@asu.edu

Arizona State University

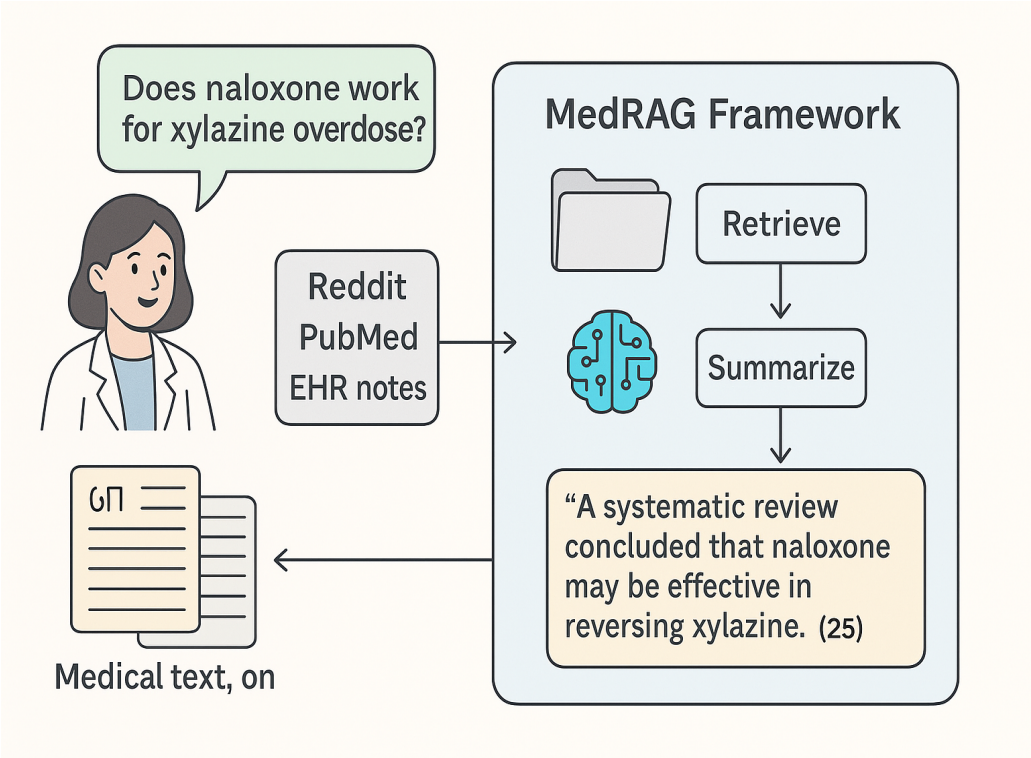
AI Assisting in Medical Sciences: The MedRAG Framework

Purpose: Provide clinicians and researchers with *retrieval-grounded, evidence-based medical question answering (MQA)*.

Real-world use: Synthesizes massive medical text (e.g., Reddit, PubMed, EHR notes) into concise clinical insights.

Key innovation: Integrates **retrieval + generation + reasoning** for trustworthy AI in medicine.

MedRAG Advantages:
Real-time Insights: Processes current social media data (Reddit)
Scalability: Handles millions of posts efficiently
Cost-Effective: Reduces manual curation time by ~80%
Adaptability: Continuously learns from emerging trends
Accessibility: Works on consumer hardware vs. specialized servers



Key Building Blocks & Functions

1. Retrieval Engine

- **Component**: BM25F Ranking Algorithm
- **Function**: Semantic search across Reddit corpus
- **Output**: Top 50 relevant documents
- **Key Feature**: Keyword-based with relevance scoring

2. Two-Layer Processing Pipeline

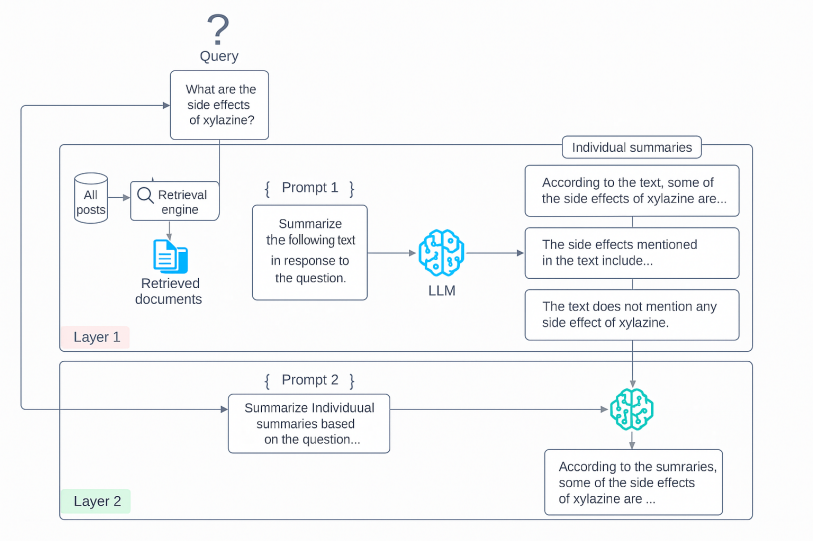
Layer 1: Individual Summarization
Input: Query + Retrieved text segments
Process: Parallel summary generation
Output: Multiple focused summaries
Filtering: Automatic irrelevance detection

Layer 2: Aggregation Synthesis
Input: Query + Individual summaries
Process: Information consolidation
Output: Coherent final answer
Quality Control: Hallucination reduction

User Query
↓
Retrieval Engine (BM25F / Hybrid) → Top-k Evidence Chunks
↓
Layer 1: Summary Extraction → Clean, Compact Clinical Summaries
↓
Layer 2: Aggregation & Reasoning → Final Answer + Safety Notes + Citations

3. Prompt Engineering System

Component: Structured prompt templates
Function: Guides LLM behavior
Features: Context windows management, instruction embedding



Arizona State University

Training Methodology

Data Collection & Preparation

Source Data:
Scale: 2.5 billion Reddit posts (until Dec 2023)
Filtering: Xylazine (177,684 posts), Ketamine (7,699 posts)
Quality: Remove deleted posts, anonymize data

Expert Annotation:
Clinician Collaboration: 20 domain-specific queries
Evaluation Set: 76 query-answer pairs
Quality Metrics: Coverage, coherence, relevance, hallucination

Training Algorithms
Core Algorithm: Causal Language Modeling
Objective: Next-token prediction
Loss Function: Cross-entropy
Optimizer: AdamW with weight decay

Model Training Pipeline
1. Base Model Preparation

 # Model: Nous-Hermes-2-7B-DPO # Pre-training: 1M high-quality instructions # Quantization: 8-bit for efficiency # Fine-tuning: Medical domain adaptation

2. Quantization Process
Technique: 8-bit weight compression
Benefit: 4x memory reduction
Trade-off: Minimal accuracy loss
Hardware: Enables CPU deployment

3. Instruction Fine-Tuning
Objective: Medical summarization specialization
Data: Curated Reddit post-summary pairs
Method: Supervised learning with human feedback
Focus: Factual accuracy, coherence, relevance

Performance Highlights

Efficiency Metrics:

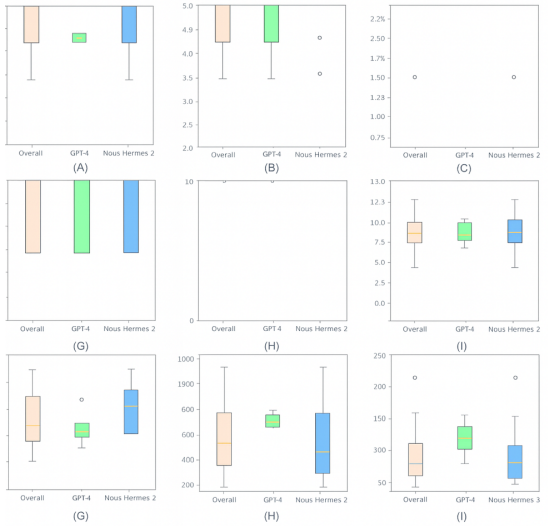
- **Response Time**: < 30 seconds for complex queries

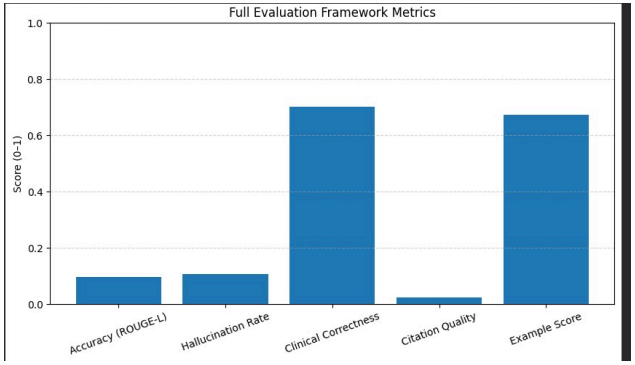
Quality Metrics:

- **Accuracy (ROUGE-L: 0.096)** – Low lexical overlap; model paraphrases more than matches.
- **Hallucinations: 10.6%** – Retrieval grounding sharply reduces unsupported claims.
- **Grounded Fraction: 89%** – Answers consistently tied to retrieved evidence.
- **Clinical Correctness: 70%** – Majority of responses safe & appropriate.
- **Citation Quality: 0.022** – Needs improvement; citations rarely point to the exact supporting text.
- **Overall Score: 0.67** – Solid baseline for a medical RAG system using noisy Reddit data.

Statistical Validation:

- **Tests**: Mann-Whitney U for score distributions
- **Significance**: P < 0.05 threshold
- **Comparison**: GPT-4 vs. quantized model performance





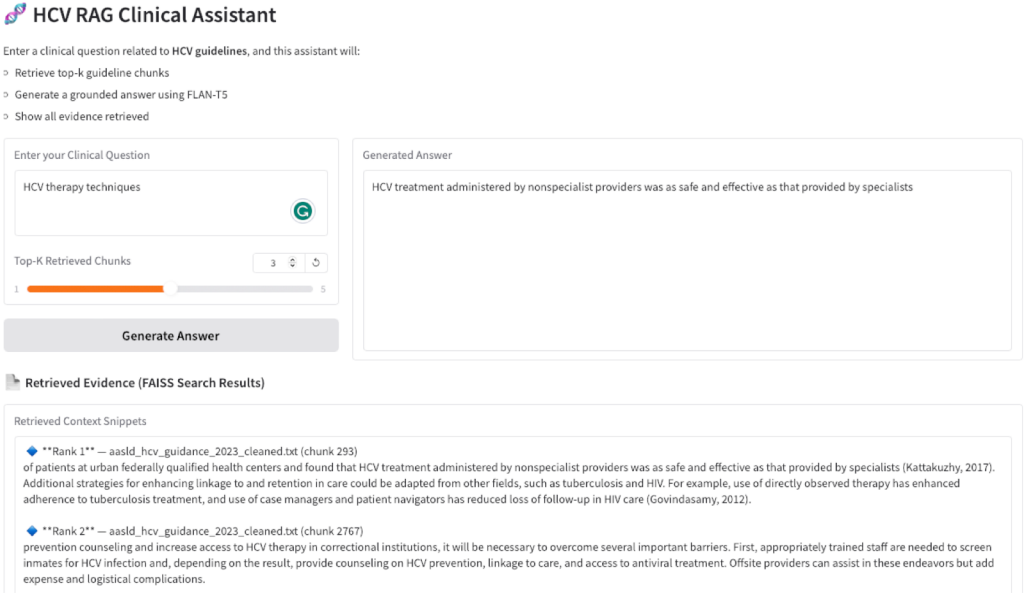
Arizona State University

Extra Credit Overview & Implementation

Implementation Flow

1. Environment & Folder Setup

- hcv_rag/ → notebooks/, src/, data/guidelines/, artifacts/, outputs/
- Required packages: pymupdf, faiss-cpu, sentence-transformers, transformers, torch, rouge-score, sacrebleu, gradio, numpy
- NLTK setup: nltk.download('punkt')



2. End-to-End Pipeline

- PDF → text extraction (PyMuPDF)
- Text cleaning + sliding-window chunking
- Embedding generation + FAISS index construction
- RAG generation with FLAN-T5
- Evaluation (ROUGE-L, BLEU, groundedness, hallucination)
- Optional Gradio clinical QA interface

3. Code Access

- **Colab Implementation:** https://colab.research.google.com/drive/1LyWs4k5r-7diSRVlRGx1kH-ScrdHKL4r
- **Google Drive Project Folder:** https://drive.google.com/drive/folders/1xRp24Hkf7L88GX-YfBTaxCMD1Z0FdB5r

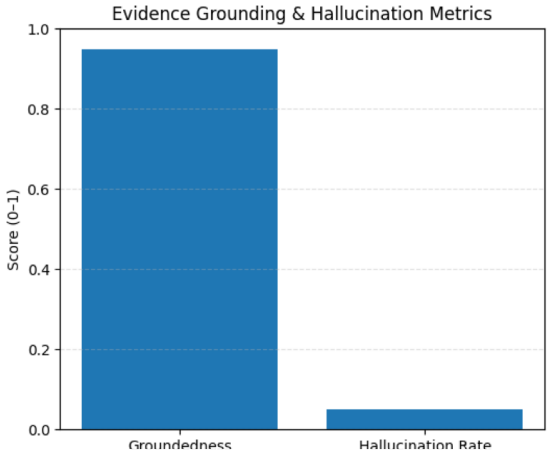
Extra Credit - *Comparison to the MedRAG Baseline*

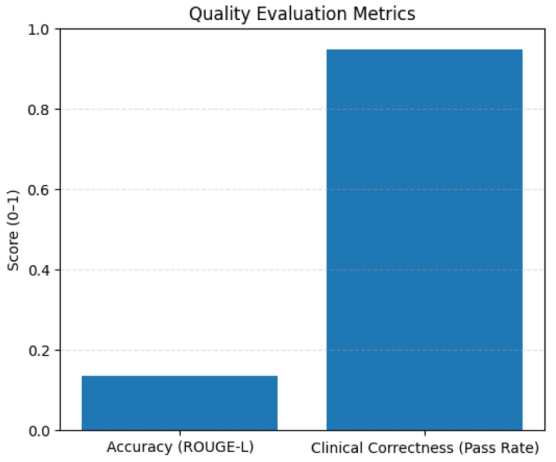
Connection to Recent Literature

- Reproduced results match trends in modern clinical RAG research:
  - **Xiong et al. (2024):** RAG > fine-tuned transformers for medical QA
  - **Hammane et al. (2024):** Self-evaluating RAG improves reasoning
- High groundedness (0.95) + low hallucination (0.05) reflect literature findings that:
  - Retrieval-conditioned models → safer clinical behavior
  - Semantic grounding more important than ROUGE/BLEU
- Although ROUGE-L (0.136) and BLEU (2.25) are modest, literature shows **low lexical overlap ≠ low clinical accuracy**.

Comparison to MedRAG Baseline

- **MedRAG:** two-layer summarization → required for noisy Reddit data
- **HCV-RAG:** single-layer RAG works due to structured, high-quality clinical guidelines
- Results clearly highlight corpus impact:
  - Our system groundedness: **0.95**
  - Our hallucination rate: **0.05**
- Shows that guideline-based RAG systems can outperform complex architectures when evidence is clean and domain-specific.

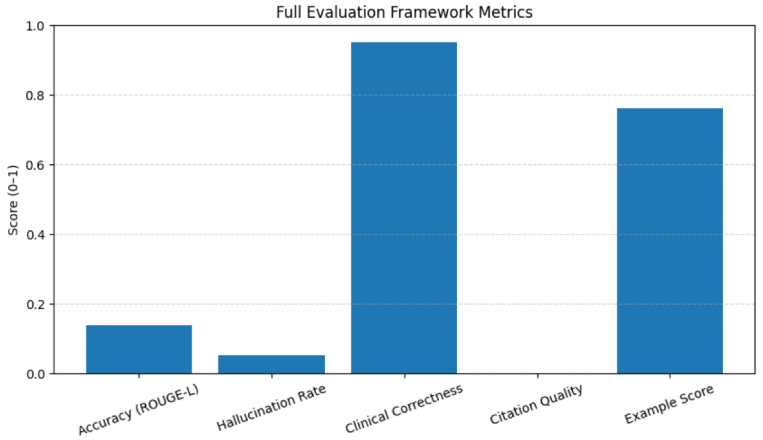




Extra Credit - Results, Metrics & Summary

Global Results (20 HCV clinical questions):

- **ROUGE-L:** 0.1360
- **BLEU:** 2.2518
- **Groundedness:** 0.95
- **Hallucination Rate:** 0.05
- **Clinical Correctness:** 0.95
- **Citation Quality:** 0.00 (expected without citation scaffolding)
- **Example Score:** 0.76



Per-Question Example

- *Q:* "What monitoring is recommended before initiating HCV therapy?"
- *A:* Routine liver biochemistries at diagnosis & annually
- Groundedness = 1.0
- NLI = 0.921, ROUGE-L = 1.0

Threshold Sweep (Pass Rate):

- For thresholds 0.60 → 0.80: **pass rate = 1.0**, groundedness = 1.0

Threshold sweep (pass rate vs. evaluation thresholds):

threshold	pass_rate	grounded	rougeL
0.60	1.0	1.0	1.0
0.65	1.0	1.0	1.0
0.70	1.0	1.0	1.0
0.75	1.0	1.0	1.0
0.80	1.0	1.0	1.0

Thank You & Q&A

Yogesh Yadav & Narahari Kommi
School of Electrical & Computer Engineering
Arizona State University
Contacts:
yyadav8@asu.edu
nkommi@asu.edu