Medical RAG Assistant — Project Report — Gehna Ahuja

# Requirement 1: Create or source three different types of data (documentation, forums, blogs)

* **Clinical Guidelines (Docs):**  
  We curated synthetic guidelines on **Diabetes, Hypertension, Asthma, CKD, and Obesity**. Each guideline mimics real-world structure with sections on diagnosis, lifestyle, and medications.  
  *Example:* “First-line for Type 2 Diabetes is Metformin. Add GLP-1 agonist if uncontrolled after 3 months.”
* **Patient Forums:**  
  We simulated JSONL threads with OP (original post), replies, and accepted answers. Topics included **Metformin side effects, white coat hypertension, rescue inhalers, CKD medications, obesity treatments, NSAID kidney risks**.
* **Medical Blogs:**  
  Blogs summarized **latest treatments** and trends. Example entries:
  + “GLP-1 receptor agonists reduce HbA1c and cause weight loss.”
  + “SGLT2 inhibitors slow CKD progression.”
  + “2025 trends in obesity: GLP-1 and CV benefits.”
* This gave us **heterogeneous sources** reflecting formal, informal, and research-driven content.

# Requirement 2: Implement a chunking strategy appropriate for each data source

* **Docs:** Split by **headings + 2 paragraphs** → ensures chunks are long enough to carry context but short enough for retrieval.
* **Forums:** Bundled **OP + top replies** → preserves discussion flow.
* **Blogs:** Split by **H2/H3 + paragraphs** → aligns with topical blog sections.
* Result: Each chunk is **coherent, self-contained, and retrievable**.

# Requirement 3: Build a retrieval system that can intelligently weigh and combine results from all sources

Used a **Hybrid Retriever**:

* **BM25** (keyword exact match)
* **Dense Embeddings** (semantic match with MiniLM)

**Source Weights:**

* Docs: +0.15 (trusted)
* Blogs: +0.05 (semi-trusted)
* Forums: baseline (no boost)

# Requirement 4: Implement a reranking mechanism to improve relevance

Used **TinyBERT cross-encoder** to rerank top candidates.

Example: Query *“Are beta blockers first-line for hypertension?”* →

* Retriever may find forum posts first.
* Reranker boosts **guideline chunks** (ACEi/ARB/CCB/Thiazides) above forum noise.

# Requirement 5: Design a mechanism to handle contradictions between sourcesTested on 10 synthetic queries (diabetes, hypertension, asthma).

* Used **Bart-MNLI** for contradiction classification.
* Added **rule-based negation detection** (catch “not / never / contraindicated”).
* Resolution Policy:
  + Docs override Blogs
  + Blogs override Forums
* Contradictions are **not hidden** — both views logged and optionally displayed.

# Requirement 6: Include logging to track which sources are being used for each response

* Every query creates a **JSON log** containing:

A screenshot of a computer screen

AI-generated content may be incorrect.

* Logs allow **auditability** for each answer.

# 7. Future Work

- Add multi-turn conversation memory.

- Expand dataset with PubMed abstracts and official guidelines.

- Deploy securely on Streamlit Cloud or via ngrok.

- Add lightweight LLM generator for fluent answers.

# 8. Conclusion

The Medical RAG Assistant demonstrates how Retrieval-Augmented Generation can:

- Bridge formal medical knowledge and real-world experience.

- Provide useful, explainable, and safe answers.

- Lay the foundation for clinical decision support tools.

⚠️ Disclaimer: This system does not replace professional medical advice.

A screenshot of a computer

AI-generated content may be incorrect.

