**Medical Chatbot with Retrieval-Augmented Generation (RAG)**

**MSDS692 – Data Science Practicum II**  
Weekly Progress Reports  
**Jonish Bishwakarma**

**Week 5 – Evaluation Framework Development**

**Planned Work for the Week**

Create evaluation script and test RAG vs. LLM-only answers.

**Project Timeline**

* Week 1 – Project definition and proposal submission (DONE)
* Week 2 – Proposal revision and evaluation design (DONE)
* Week 3 – Collect medical documents and build knowledge base (DONE)
* Week 4 – Implement retrieval pipeline and integrate with LLM (DONE)
* Week 5 – Develop evaluation script and run baseline testing (DONE)
* Week 6 – Debugging and refinement
* Week 7 – Build chatbot interface
* Week 8 – Final evaluation, analysis, and presentation

**Progress for the Week**

* Implemented evaluation script (evaluate.py).
* Designed JSON file of test questions (mix of medical and out-of-scope).
* Wrote functions:
  + rag\_answer() – retrieves docs and answers.
  + llm\_only\_answer() – baseline model without retrieval.
  + compute\_scores() – calculates accuracy, faithfulness, and privacy.
* Ran small batch of test questions and saved results to CSV.

**Roadblocks/Issues**

Some “in-scope” questions were marked incorrect when RAG responded “I don’t know.” This lowered RAG’s accuracy compared to LLM.

**Plan for Next Week**

Refine evaluation logic, debug issues, and improve faithfulness measurement.

**Resources for the Week**

* Guha, M. (2007). *The Gale Encyclopedia of Medicine (3rd edition).* Detroit: Thomson Gale.