**Model Refinement**

1. **Overview**

The model refinement phase focuses on optimizing the RAG-based CDSS to enhance its performance in retrieving relevant medical knowledge from 17 annotated medical textbooks and generating accurate, evidence-based clinical recommendations. This phase is critical to improving the system’s retrieval accuracy, response quality, and computational efficiency, ensuring alignment with the project’s goals of reducing misdiagnoses and supporting Sustainable Development Goals (SDGs) like SDG 3 (Good Health and Well-being). Refinement involves iterative adjustments to the retrieval and generation components, guided by evaluation metrics and clinical relevance.

1. Model Evaluation

In the initial model exploration (referencing the prototype in cdss\_rag\_prototype.py), the RAG system used a sentence transformer (all-MiniLM-L6-v2) for retrieval and BART (facebook/bart-large) for generation. Key metrics from the exploration phase included:

* Retrieval Accuracy: Precision (proportion of top-5 retrieved chunks relevant to the query) was approximately 70% on a small set of clinical queries (e.g., “What are the treatment options for hypertension?”).
* Generation Quality: Evaluated qualitatively, BART generated coherent responses but occasionally included redundant information or lacked specificity (e.g., generic treatment suggestions).
* Latency: Retrieval and generation took ~5–7 seconds per query on a single GPU, indicating Areas for Improvement:

1. Refinement Techniques

* Testing on Gemini: **Gemini 2.5 models are capable of reasoning through their thoughts** before responding, resulting in enhanced performance and improved accuracy.
* Chunk Optimization: Reduced chunk size from ~300 to ~200 words to increase retrieval granularity, ensuring more precise matches.

These techniques aimed to improve retrieval accuracy, response quality, and system efficiency, critical for real-time clinical decision support.

1. **Cross-Validation**

**Initial Strategy**: The initial model used a simple train-test split (80:20) on a small set of clinical queries and textbooks for evaluation.

**Changes Made:**

* Created a validation dataset and clinical queries derived from medical case studies and textbook-based questions, ensuring coverage of all 17 textbooks.
* Stratified the folds by medical discipline (e.g., cardiology, neurology) to account for domain-specific variations.

1. Feature Selection:   
   Embedding Dimensionality: the number of dimensions or values used to represent data in a transformed numerical space. Essentially, it's the length of the vector that represents an object  
     
   **Impact on Performance:** Higher dimensionality allows for more complex and nuanced representations of data, but requires more computational resources. Lower dimensionality simplifies calculations but may oversimplify patterns**.**

**Test Submission**

* 1. **Overview**

The test submission phase involves preparing the refined RAG model for evaluation on a test dataset of clinical queries, simulating real-world clinical scenarios. This phase ensures the model is ready for deployment by validating its performance on unseen data, measuring key metrics, and preparing for integration into a clinical environment. Steps include finalizing the model, preparing the test dataset, applying the model, and evaluating results.

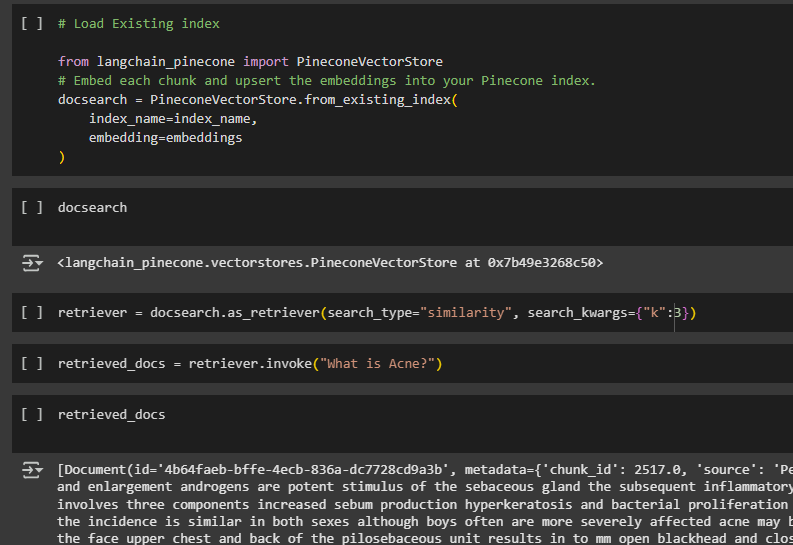
* 1. **Data Preparation for Testing**

Queries cover diverse topics (e.g., “What are the diagnostic criteria for type 2 diabetes?” or “What surgical approaches are recommended for appendicitis?”) and are balanced across the 17 textbooks’ disciplines. Each query is paired with ground-truth answers (e.g., textbook excerpts or expert-validated responses) for evaluation.

* 1. Model Application

The refined RAG model was applied to the test dataset as follows:

* **Retrieval**: Each query was encoded using the fine-tuned sentence transformer, and the top-5 relevant chunks were retrieved from the FAISS index.
* **Generation**: Retrieved chunks were combined with the query and fed into Gemini 2.5 Flash model to generate a clinical recommendation.



* 1. Test Metrics
* Retrieval: Precision (proportion of top-5 retrieved chunks relevant to the query)
* Generation: for text similarity to ground-truth answers
* for long-form response quality, and clinical relevance (manually assessed by comparing responses to expert-validated answers).
* Latency: Average time to process a query (retrieval + generation).
  1. Model Deployment  
     User Interface: Developed a simple web interface for clinicians to input queries and view responses, including retrieved textbook references for transparency.