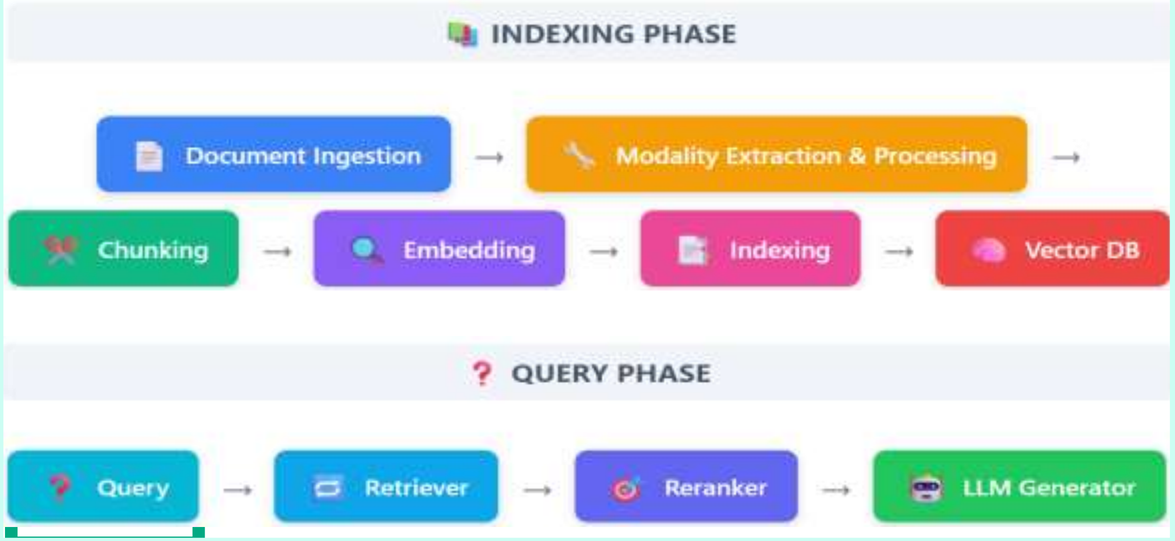
Project Details

Title: Use case driven multimodal RAG pipeline recommender

Duration: Feb 2026 to Oct 2026

Skills: GenAI, multimodal RAG, parsing and extraction, NVIDIA NIMs, RAG evaluation, containers, optimization

Sample RAG Pipeline



Project Description

- Multimodal RAG framework is becoming more flexible and modular to support multiple registries, models and libraries in the same pipeline
    - E.g. Models from Ollama, NVIDIA NIMs, or libraries etc.
- Now customers can choose different models at different phases of a RAG pipeline based on –
    - Use case (finance, legal etc.)
    - System capability (GPU, RAM etc.)
    - Available models / licenses
- For a specific use case and HW resources, develop a multimodal RAG recommender system that can –
    - Evaluate all possible pipelines that can be built with different supported models
    - Suggest the best optimal pipeline as per user need

Project Deliverables

Phase-1 (Duration = 2 months)

1. Build a modular & flexible multimodal RAG framework for PDF processing and test it
   a) PDF processing using 2 separate containerized pipelines – one for "ingestion+embeddings" & another for "retrieval+generation"
   b) Seamless integration with different repositories (Ollama, NVIDIA NIM, vLLM etc.) for model serving. Configuration based easy switching from one model to another across repositories. Each pipeline can support different models across repositories
2. Output = Design, code, and test results

Phase-3 (Duration = 2 months)

1. Develop a multimodal RAG evaluation framework to capture different performance metrics
   a) Either build a new framework or use existing such as RAG-Check, LlamaIndex etc. Framework must be containerized
   b) Define metrics to be captured –
      - Quality related such as retriever score, generator score, precision, recall, faithfulness etc.
      - Performance related such as latency, throughput, Time to First Token (TTFT), Tokens Per Second (TPS) etc.
   c) Use or develop multiple benchmark / curated data sets representing different use cases (e.g. - finance, legal etc.)
   d) Evaluate RAG pipelines for a given use case and capture metrics
2. Output = Design, code, and test results

Phase-2 (Duration = 2 months)

1. Build a modular & flexible multimodal RAG framework for Video processing and test it
   a) Video processing using 2 separate containerized pipelines – one for "ingestion+embeddings" & another for "retrieval+generation"
   b) Seamless integration with different repositories (Ollama, NVIDIA NIM, vLLM etc.) for model serving. Configuration based easy switching from one model to another across repositories. Each pipeline can support different models across repositories
2. Output = Design, code, and test results

Phase-4 (Duration = 2 month)

1. Build a simple recommender system that can suggest the optimal RAG pipeline for a given use case and HW resources
   a) Get all available models (from all supported repositories)
   b) Get user objective at a high level (e.g. *Get me a performance optimized / highly-accurate / low-cost RAG pipeline*)
   c) Run the use case with multiple possible pipelines that can be built using different available models and capture all metrics
   d) Map user objective to collected metrics to get optimal pipeline
2. Output = Design, code, and test results

Phase-5 (Duration = 1 month)

1. Publish and present code, report, result, and slides