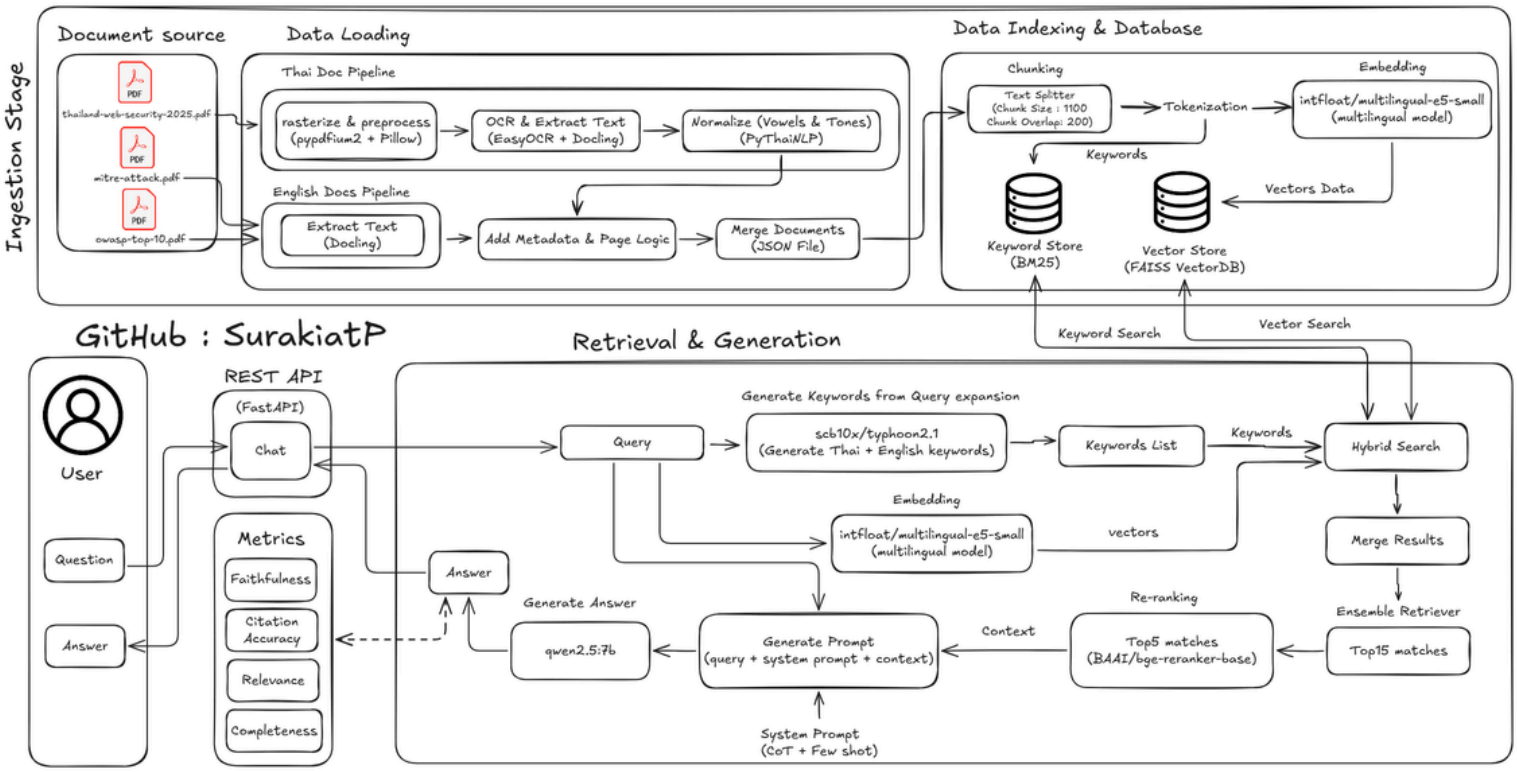
CYBER RAG PIPELINE SYSTEM ARCHITECTURE

Retrieval-Augmented Generation for Cybersecurity Document QA

[Question] → [Query Expansion] → [Retrieval] → [Reranking] → [Generation] → [Answer + Citations]



Project Overview

- RAG system designed to answer cybersecurity questions using three authoritative documents:
  → mitre-attack-philosophy-2020.pdf, owasp-top-10.pdf and thailand-web-security-standard-2025.pdf.
- Built on hybrid retrieval architecture combining keyword-based (BM25) and semantic search (FAISS), enhanced with query expansion and reranking to maximize retrieval precision.
- All responses are strictly grounded in source documents with explicit citations, ensuring zero hallucination and full traceability.

Core Capabilities

- **Strict Grounding**
  All answers derived exclusively from the 3 provided PDFs
- **Precise Citations**
  Every claim includes document name + page number reference
- **Zero External Data**
  No internet search, no general knowledge injection
- **Hallucination-Free**
  LLM constrained by retrieval context only

Advanced Techniques

- **Query Expansion**
  SCB10x/Typhoon2.1 generates Thai+English keywords
- **Hybrid Retrieval**
  BM25 (keyword) + FAISS (semantic) → Top-15 chunks
- **Reranking**
  BAAI/bge-reranker-base refines to Top-5 most relevant
- **Evaluation Framework**
  Manual scoring: Faithfulness, Citation, Relevance, Completeness

Architecture Design & Decision Rationale

- **Local Deployment (Ollama)**
  Qwen2.5:7B-Q4 for generation, Typhoon2.1 for expansion → Full control, no API costs, privacy compliance
- **Multilingual Embedding (intfloat/multilingual-e5-small)**
  Handles Thai+English PDFs seamlessly → Single model for all 3 documents
- **Hybrid Retrieval (BM25 + FAISS)**
  BM25: Exact keyword matching (e.g., "OWASP") ,FAISS: Semantic understanding (e.g., "access control breach" → vector)
  → Complementary strengths, higher recall
- **Reranking Layer (BAAI/bge-reranker-base)**
  Cross-encoder rescores Top-15 → Top-5 → Precision boost without retrieval overhead

Metric (Evaluated on 10 test queries) → (*EVALUATION.md*)

Faithfulness → 91.20%, Citation Accuracy → 85.00%, Relevance → 4.20/5.00, Completeness → 4.10/5.00

GitHub Repository → ***https://github.com/SurakiatP/cyber-rag-assignment***

Surakiat Kansa-ard