RAG Pipeline V2

# Problem

Organizations keep critical knowledge scattered across PDFs, Word, Excel, and images. Finding precise answers often requires multi‑step reasoning across documents (e.g., extracting numbers from models and reconciling with narrative). Manual search is slow, but even then, many Accounting, Legal, and Financial departments in general spend a lot of time in extracting key information. It is a time-consuming task which takes almost whole day for some employees

We need a reliable, citation‑grounded RAG system that works across mixed formats used widely across these departments and handle noisy metadata.

# Track: RAG & Knowledge Mastery

Reduce hallucinations, improve retrieval and reranking—e.g., personal productivity assistants that learn your workflows. We focused especially on Hallucinations, retrieval, and numerical accuracy. Obtained 85% MRR and 100% Relative Numerical Accuracy.

# Stack

LlamaParse for robust parsing of PDFs/Office/images. Logical documents are formed with the help of Custom You.com agents (page‑continue checks, doctype classification, query routing). Made over 1700 API Calls during the project.

Applied Recursive character splitter followed by semantic splitting for cohesion.

Used FAISS for dense vector search + BM25 for lexical search. Late fusion combines both; a cross‑encoder reranker promotes the best passages. Gemini 2.5 Flash handles generation.

SubQuestionQueryEngine decomposes complex queries and routes sub‑queries to named tools (per‑doctype engines and a global engine). Soft doctype preference can be enabled, but a no‑metadata global tool fallback is available when labels are noisy.

Notebook and prebuilt logical documents (pickle) have been provided at [github](https://github.com/anubhav-ds/You.com-Hackathon-RAG/tree/main) to enable reviewers run everything without external LlamaParse parsing and running you.com API Calls.