Educational RAG pipeline

Introduction & Project Goal

PROBLEM

Students need a personal assistant for their study materials.

SOLUTION

A Retrieval-Augmented Generation (RAG) pipeline:

- Leverages open-source embedding & generative AI models
- User-friendly Gradio interface

CORE GOAL

Facilitate an effective Human-AI Interaction experience, emphasizing usability, interpretability, and trust.

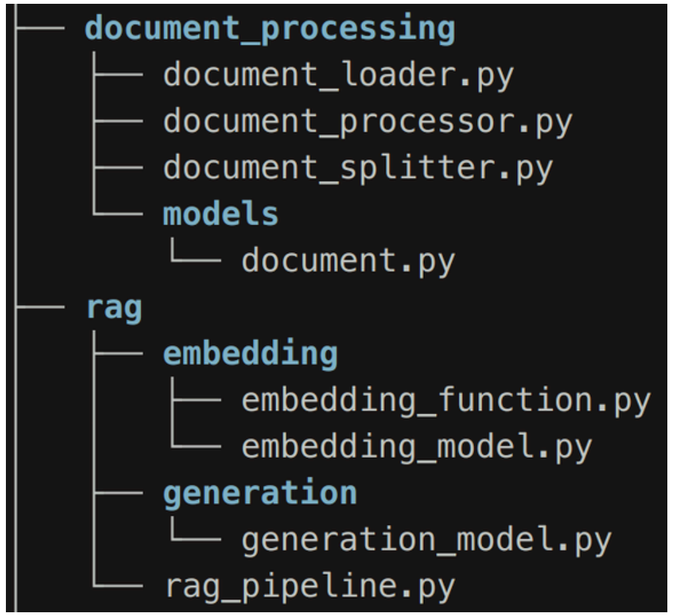
ALIGNMENT

Meets course requirements for an ML system demonstrating human-centric design.

System Architecture Overview

KEY COMPONENTS

- Embedding Model
- Generation Model
- Vector Database
- Document Processing Pipeline
- RAG Orchestrator
- User Interface



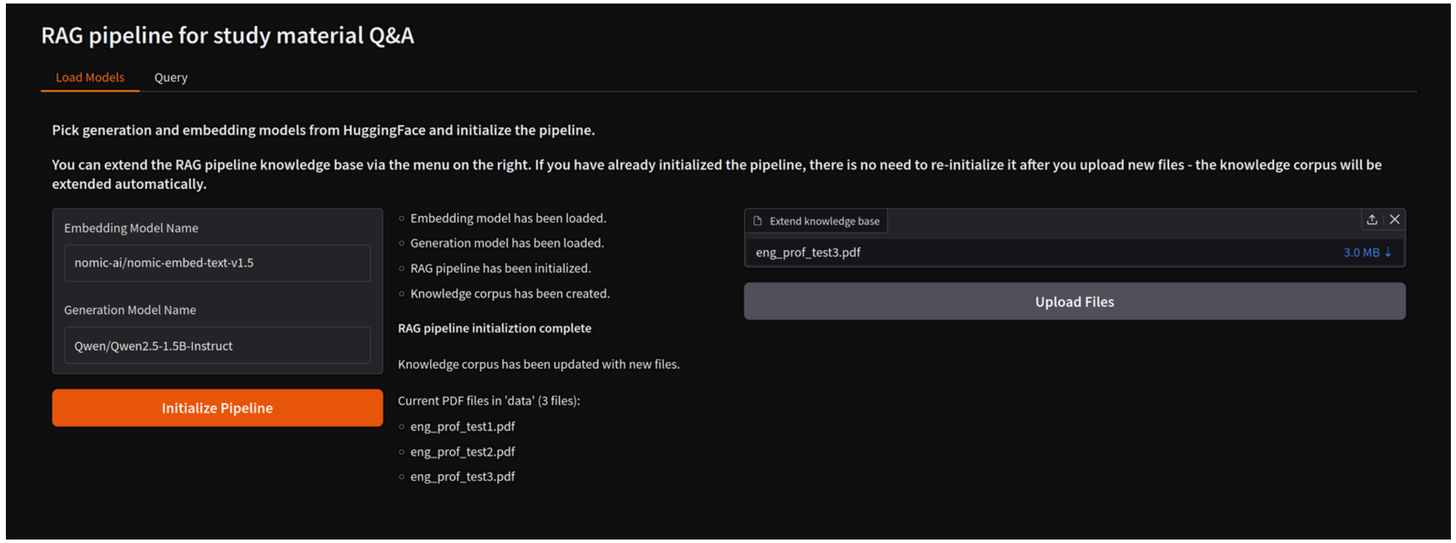
UI Design
"Load Models" Tab

PURPOSE

Initial pipeline setup & knowledge base extension.

KEY DESIGN CHOICES & HAIID PRINCIPLES

- Model Selection: User control & freedom (Nielsen, HAX G17)
- "Initialize Pipeline" Button & Status: Clear intention, step-by-step feedback, visibility of system status (Norman, Nielsen, Shneiderman)
- Corpus File Display: Transparency of knowledge base
- File Upload: User control over corpus, immediate feedback



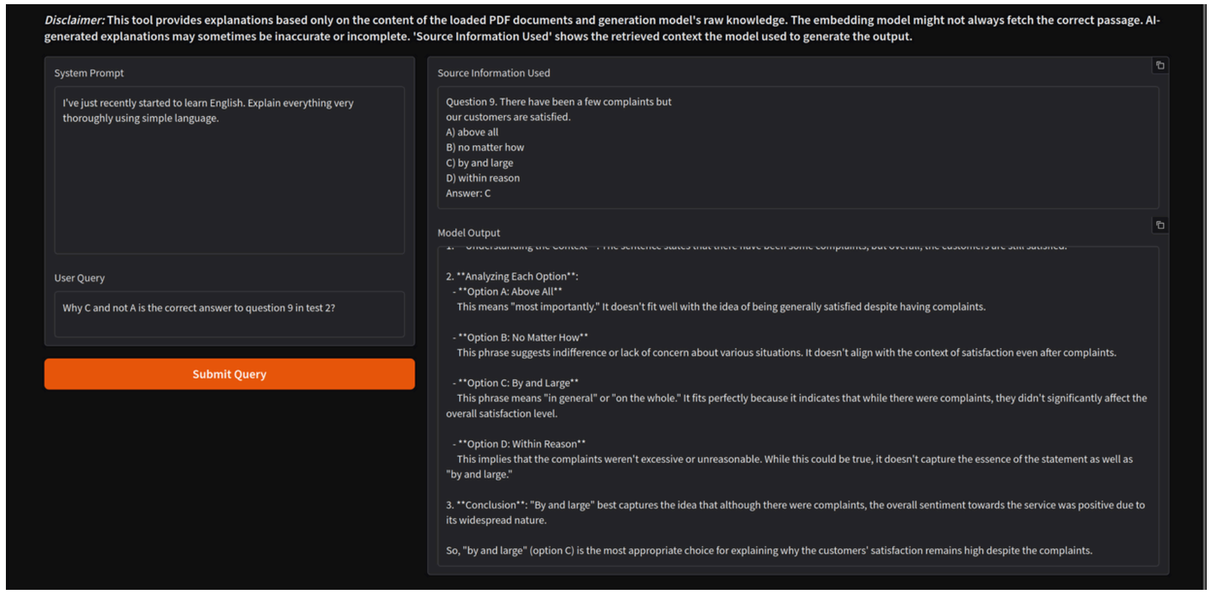
UI Design
"Query" Tab

PURPOSE

Initial pipeline setup & knowledge base extension.

KEY DESIGN CHOICES & HAIID PRINCIPLES

- Disclaimer: Sets right expectations (Google People+AI P2, HAX G2), ethical consideration
- System Prompt: User control (Nielsen, HAX G17), enables social nature of explanations (Miller)
- "Source Information Used" Textbox: critical for interpretability & transparency (HAX G11), selective explanation
- "Model Output" Texbox: streaming provides continuous feedback (Norman, Shneiderman), enhaces UX



Key HAIID Principles Applied

TRANSPARENCY & INTERPRETABILITY

- "Source Information Used" is paramount
- Display of model names, corpus files

USER CONTROL & FREEDOM

- Model selection, corpus management (uploads), query formulation, system prompts

FEEDBACK & VISIBILITY OF SYSTEM STATUS

- Status messages (initialization, uploads), streaming output, error messages

Key HAIID Principles Applied

SETTING EXPECTATIONS & ACCOUNTABILITY

- UI Disclaimer on AI limitations
- Source context allows users to check the factual basis of the AI's answers

HUMAN-CENTERED DESIGN

- Solves specific user need (Q&A for studying)
- Augments user ability, user is in control

ETHICAL CONSIDERATIONS

- Disclaimer, transparency features
- Local deployment respects data privacy
- RAG approach can reduce reliance on general LLM bias by grounding in knowledge corpus

TEXTUAL DISPLAY

- “Source Information Used”: Direct representation of retrieved context
- “Model Output”: Dynamic textual visualization via streaming
- Pipeline State/Corpus Files: Status & list formats

LAYOUT & GROUPING

- Tabs visually separate interaction stages
- Grouping aids recognition rather than recall

IMPLICIT PROCESS VISUALIZATION

- Interaction sequences (e.g., file upload -> list update) communicate internal processes
- "Source Information Used" supports interpretability, a key InfoViz goal in Human-Centered AI

Information Visualization Techniques

Lessons Learned

EXPLAINABILITY IS CORNERSTONE OF TRUST IN RAG

- Simply providing an answer is insufficient. We need to turn black box into a verifiable assistant

PROACTIVE EXPECTATION MANAGEMENT IS CRUCIAL

- Upfront disclaimer on accuracy/boundaries is vital

CLEAR VISIBILITY & USER CONTROL MITIGATE AI OPACITY

- Continuous feedback & user control over inputs improve usability

ITERATIVE DESIGN IS KEY FOR EFFECTIVE HAIID IMPLEMENTATION

- Translating abstract principles to concrete UI requires refinement