Clinton Iwu
4/28/25
ITSC-4010
Mini Project 2

Report

Overview

Agent 1 was developed as an intelligent legal assistant capable of answering user-uploaded legal questions by reading PDFs and applying advanced Retrieval-Augmented Generation (RAG) techniques. It utilizes a combination of LLM reasoning and vector database retrieval to ensure precise, context-aware responses.

Implemented Techniques

PDF Document Ingestion

The process begins when a user uploads a legal PDF document. The unstructured library is used to parse and extract raw text from the PDF while preserving the document's structure and content integrity for downstream processing. The extracted text is then segmented into semantically meaningful chunks using the RecursiveCharacterTextSplitter.

Embedding and Vector Storage

Each text chunk is embedded using the HuggingFaceEmbeddings module with the all-mpnet-base-v2 model. These embeddings are stored in ChromaDB, a vector database that allows efficient semantic retrieval based on user queries.

Query Rewriting

User-submitted queries are often vague or lack the necessary keywords to yield relevant results. To address this, a query rewriting module powered by LLaMA 3 via Groq reformulates the original query into a clearer and more specific legal question. This enhancement improves the precision of document retrieval.

HyDE (Speculative RAG)

HyDE, a speculative RAG technique, is used to generate a hypothetical answer before retrieving documents. This generated answer is then embedded and used as the actual search vector. This approach enhances recall and enables more contextually relevant document matches.

Example Output and Explanation Original Query: “Can I use someone else's idea and trademark it under my company name?”

The system first rewrites this question to make it more specific and legally relevant:

> Rewritten Query: “Can I use someone else's intellectual property (e.g., a concept, design, or idea) and register a trademark for a product with my company name, without obtaining permission or compensation from the original creator, and without infringing on their intellectual property rights? If so, what are the legal implications and potential consequences of doing so?”

Using this clarified query, HyDE generates a speculative answer based on the LLM’s internal legal knowledge. For example:

> Speculative Answer (HyDE Output):
>
> “Generally, using someone else's intellectual property without authorization and attempting to trademark it under your own company name may result in legal consequences, including trademark rejection, lawsuits for infringement, and potential financial penalties.”

This speculative answer is then embedded and used to retrieve the most semantically relevant sections from the uploaded legal PDF. These retrieved sections are combined with the speculative answer and fed back into the LLM to formulate the final response. As a result, the system is able to deliver a more accurate and well-contextualized legal explanation tailored to the user's question.

Agent 2 – AI-Driven Scientific Research Assistant

Agent 2 was developed as an intelligent assistant focused on scientific research. Its primary goal is to refine vague queries, generate hypothetical abstracts using speculative RAG, and rank retrieved documents to prioritize the most relevant academic sources. Built on the same architecture as Agent 1, it uses LangChain, Groq’s LLaMA model, HuggingFace embeddings, Chroma vector storage, and langchain_unstructured for PDF functionality

Refined Query Generation Agent 2 begins by refining vague or general user input into a clear and specific research question. For example, the input

> "What about medical AI in healthcare?" is transformed into: "What is the current state and projected growth trajectory of medical artificial intelligence (AI) assistants in healthcare, and what factors will influence their adoption and utilization in the next 5–10 years?"

Hypothetical Abstract and Speculative RAG

The agent uses speculative RAG to create a hypothetical abstract that resembles a real academic summary. This abstract outlines the current and future roles of medical AI assistants, identifies key adoption factors (data quality, regulation, clinician trust), and forecasts industry growth. The speculative answer is embedded and used as the primary search vector for document retrieval.

Document Retrieval and Ranking

After creating the speculative query, Agent 2 searches the Chroma vector store for matching documents. It ranks them using cosine similarity and relevance checks to find the ones that best match the refined research question and abstract. This ensures the final results are based on the most relevant and accurate sources.

Supporting Evidence

Agent 2 found and ranked documents, including a study showing RCMed works well for many medical conditions with high accuracy and flexibility. It also included important research about AI in healthcare, such as:

- Growth projections from MarketsandMarkets (to $31.4 billion by 2025)
- Research on adoption drivers like regulatory clarity, cost, and trust (e.g., JAMA, Journal of Healthcare Engineering)
- Specific validation studies from the user-uploaded research_temp.pdf involving rare conditions and diverse populations

Strengths and Human Oversight

Agent 2 is effective at creating clear research questions and using smart document ranking with speculative reasoning to improve results. This helps make its answers more accurate and complete. However, human review is still important to check for errors, ensure the information is correct, and avoid relying too much on guesses.