



OrderFox

Retrieval-Augmented Generation LLM Agent

Bayes Brigade

Harald Semmelrock, Felix Schatzl, Sebastian Brunner, Ben Ellis

Problem description

- Transform messy web data into structured knowledge and build an intelligent agent on top of it.
- Develop a working prototype of a RAG agent that can answer user questions based on a web-scraped company dataset, built with the target user being a supply chain director

Background

- Large Language Models (LLMs)
- Retrieval-Augmented Generation (RAG)
- Text Embeddings
- Vector Databases
- Inverted Index
- Prompt Engineering



(1) Parsing and Preprocessing

- Removal of non-informative .css and .js pages using regex on URLs
- Filtered out pages with _jb_static or _static in the path
- Skipped files that couldn't be decoded with UTF-8
- Reduced data from 17 GB → 5 GB after initial cleanup
- Applied MinHash on a subset to remove repetitive text (headers/footers)
- Achieved an additional 60% reduction on the subset

(2) Knowledge base design

- Vector database
 - ChromaDB
 - Used all-MiniLM-L6-v2”to generate embeddings
- Inverted Index
- Named Entity SQL

Split text data into 4096 character chunks with 2048 character overlap

SQL DB for structured info like regions/products, but limited by parsing quality vs speed

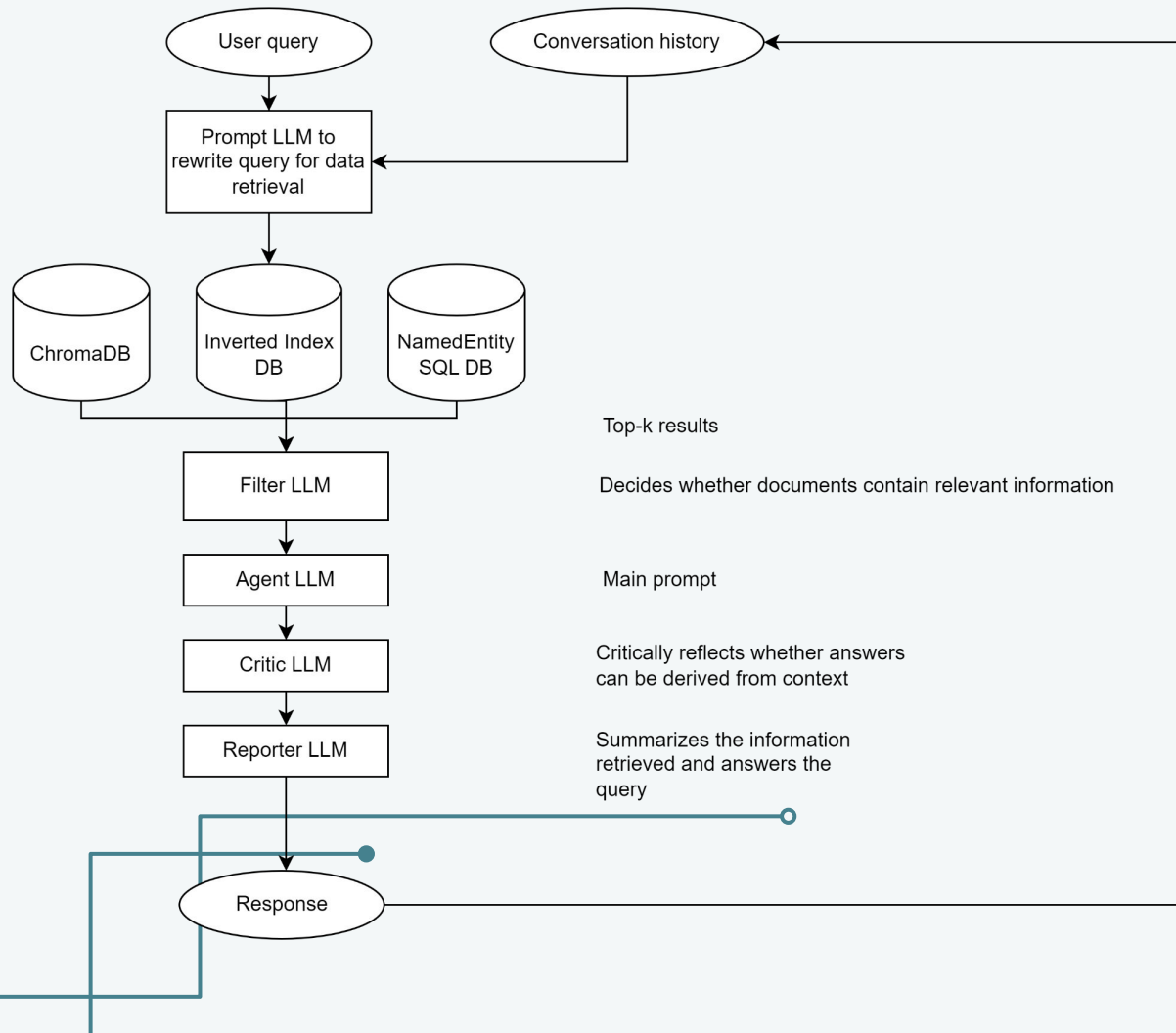
- Inverted Index added for fast keyword-based document matching
- Combined vector and inverted index for hybrid retrieval

(3) Retrieval system

- User query passed through prompt-engineered o1-mini
 - Reformats into an optimized vector DB query
 - Maintains context from prior questions
- Queries Vector, Inverted Index, and Named Entity SQL Database
- Top 5 results from each retrieved using filenames and URLs
- Results re-filtered through o1-mini for context quality
- Final context passed to the generation agent

(4) Generation agent

- Takes filtered context and applies strict or loose RAG settings
- Uses a “critic” agent to validate factual accuracy
- Final output created by a “reporter” agent for clarity and completeness
- Modular pipeline allows layered checks and fine-tuned generation



(5) Bonus

- Designed a simple user chat interface for easy interaction
- Hosted interface through a tunnel for quick access by judges
- User settings available to:
 - Enable/disable each database (Vector, Inverted Index, SQL Entity)
 - Enable/disable strict RAG generation
- Positive Results:
 - Answered example and constructed queries effectively
 - Returned clear and relevant responses in a reasonable amount of time
 - Responses were helpful and aligned with the company's HTML data

Results

User query: How can i contact American Cruise Lines?

Ground Truth example: By phone: 800-460-4518 or by mail

Strict RAG answer: You can contact American Cruise Lines by calling their toll-free number at 800-460-4518. For customer inquiries, you can also reach them via email at Inquiry@AmericanCruiseLines.com. If you have a media-related inquiry, you can email Media@AmericanCruiseLines.com.

Results

User Query: Who uses Agile Methodologies to deal with Marketing in Fort Lauderdale, FL?

Retrieved Documents: ['<https://www.starmark.com>', '<https://interexy.com>', '<http://starmark.com>', '<https://stuenterprises.com>', '<https://thinkstrategic.com>', '<http://stearnsweaver.com>', '<https://1105media.com>', '<https://stuenterprises.com>']

Ground Truth example: starmark.com (and others possible)

Loose RAG answer: Starmark, an agency based in Fort Lauderdale, FL, utilizes Agile Methodologies in their marketing practices. They emphasize a team-driven environment, fostering collaboration throughout the project lifecycle, and implement bi-weekly sprints to enhance teamwork and effectiveness.

The Tool speaks for itself (and the report)

Options

☒ Vector DB

☒ Inverted Index

☐ Named Entities

☐ Strict Mode

Gieni by Bayes Brigade

Hello! How can I assist you today?

Name a company in philly

A company in Philadelphia is ****firstPRO****, which is a staffing and recruiting firm specializing in fields such as Accounting & Finance and Information Technology.

<http://firstproinc.com>

<https://www.firstproinc.com/>

<https://www.firstproinc.com/about...>

<http://cohencpa.com/contact/phil...>

<https://pharosiq.com/contact-us/>

<https://www.gilmore-assoc.com/lo...>

Type your message...

Send



The Tool speaks for itself

<http://go.snb.li/datathon>

