

Open in app ↗



Write



★ Member-only story

Building LLM Applications: Open-Source RAG (Part 7)

Vipra Singh · [Follow](#)

27 min read · Mar 16, 2024



Learn Large Language Models (LLM) through the lens of a Retrieval Augmented Generation (RAG) Application.

Posts in this Series

1. [Introduction](#)
2. [Data Preparation](#)
3. [Sentence Transformers](#)
4. [Vector Database](#)
5. [Search & Retrieval](#)
6. [LLM](#)
7. *Open-Source RAG (This Post)*
8. [Evaluation](#)

9. Serving LLMs

10. Advanced RAG

Table of Contents

- 1. Introduction
 - 1.1. LLMs
 - 1.2. LLM Providers
 - 1.3. Vector Databases
 - 1.4. Embedding Models
 - 1.5. Orchestration Tools
 - 1.6. Quality Tuning Tools
 - 1.7. Data Tools
 - 1.8. Infrastructure
- 2. Build an LLM application from scratch
 - 2.1. Prepare the data
 - 2.2. Create the embeddings + retriever
 - 2.3. Load quantized model
 - 2.4. Setup the LLM chain
 - 2.5. Compare the results
- 3. LLM Server
- 4. Chatbot Applications
- 5. Application 1: Chat with multiple PDFs
- 6. Application 2: Chatbot with Open WebUI
- 7. Application 3: Deploy Chatbot using Docker
- Conclusion
- Credits

vishnu, read the best stories from industry leaders on Medium.

The author made this story available to Medium members only. Upgrade to instantly unlock this story plus other member-only benefits.

- ★ Access all member-only stories on Medium
- ★ Become an expert in your areas of interest
- ★ Get in-depth answers to thousands of questions about technical
- ★ Grow your career or build a new one



Marc-André Giroux
Sr. Software Developer
Netflix



Carlos Arguelles
Sr. Staff Engineer
Google



Tony Yiu
Director
Nasdaq



Brandeis Marshall
CEO
DataedX



Cassie Kozyrkov
Chief Decision Scientist
Google



Memo Akten
Asst. Professor
UCSD

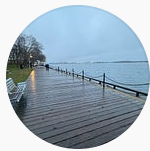


Vitali Zaidman
Software Architect
Meta



Camille Fournier
Head of Engineering
JPMorgan Chase

Upgrade



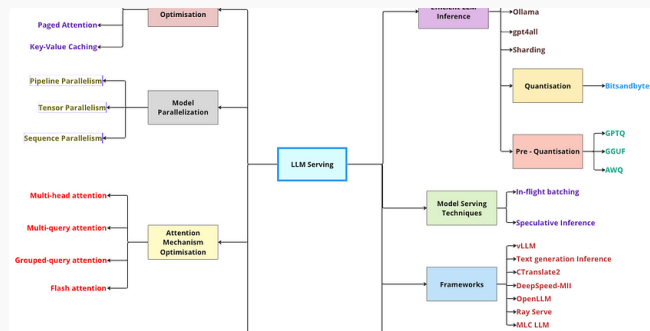
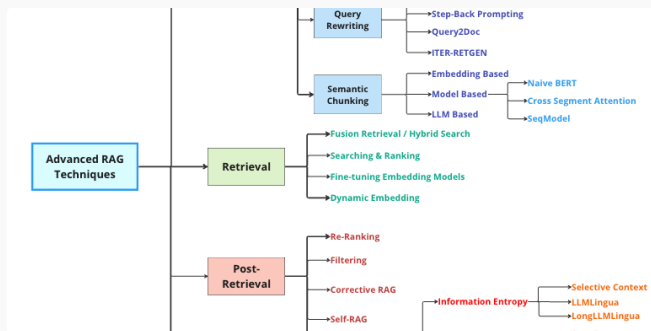
Written by Vipra Singh

1K Followers

Follow



More from Vipra Singh



Vipra Singh

Building LLM Applications: Advanced RAG (Part 10)

Learn Large Language Models (LLM) through the lens of a Retrieval Augmented...

🌟 · 48 min read · Apr 27, 2024



384



2



Vipra Singh

Building LLM Applications: Serving LLMs (Part 9)

Learn Large Language Models (LLM) through the lens of a Retrieval Augmented...

🌟 · 50 min read · Apr 17, 2024

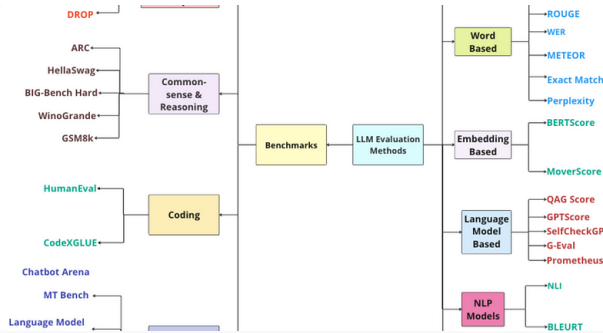


546



3





Vipra Singh

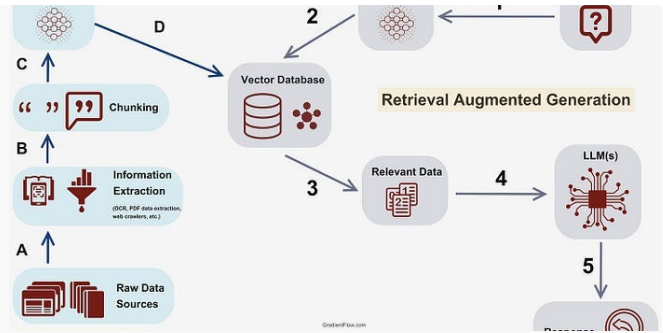
Building LLM Applications: Evaluation (Part 8)

Learn Large Language Models (LLM) through the lens of a Retrieval Augmented...

48 min read · Apr 7, 2024

276

1



Vipra Singh

Building LLM Applications: Introduction (Part 1)

Learn Large Language Models (LLM) through the lens of a Retrieval Augmented...

5 min read · Jan 8, 2024

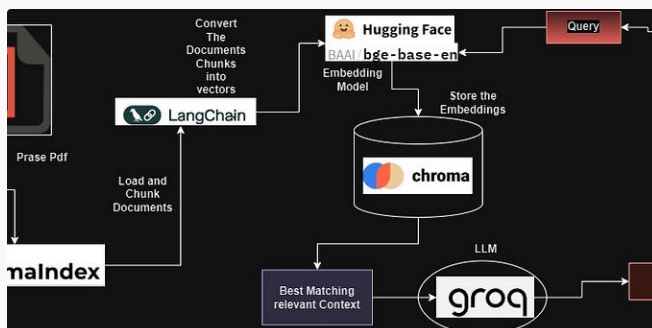
905

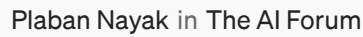
3



See all from Vipra Singh

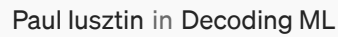
Recommended from Medium





Retrieval-Augmented Generation (RAG) is a new approach that leverages Large Language...

13 min read · Apr 7, 2024



Implement from scratch 4 advanced RAG methods to optimize your retrieval and post-...

15 min read · May 4, 2024



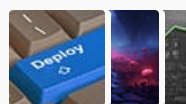
1494 stories · 1011 saves



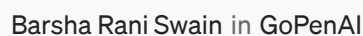
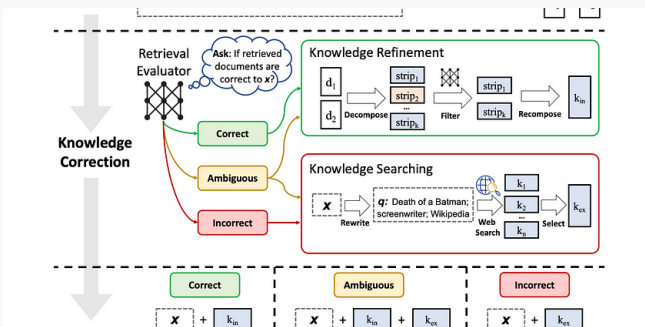
6 stories · 473 saves



47 stories · 1642 saves

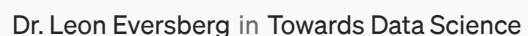
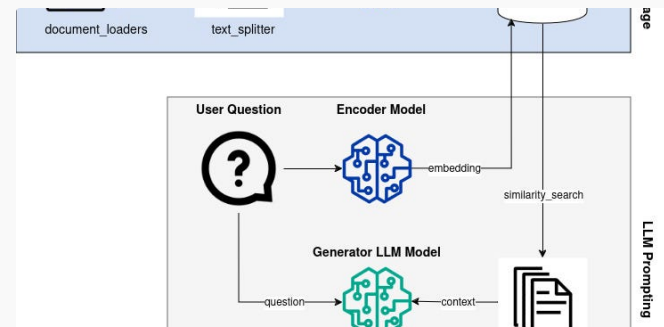


20 stories · 1254 saves



CRAG enhances the traditional RAG by introducing a retrieval evaluator to assess th...

10 min read · Apr 23, 2024



Talking to PDF documents with Google's Gemma-2b-it, LangChain, and Streamlit

✦ · 12 min read · Mar 31, 2024



320



1.4K



8



Fareed Khan in Level Up Coding

Building LLaMA 3 From Scratch with Python

Code Your Own Billion Parameter LLM

29 min read · May 28, 2024



1.1K



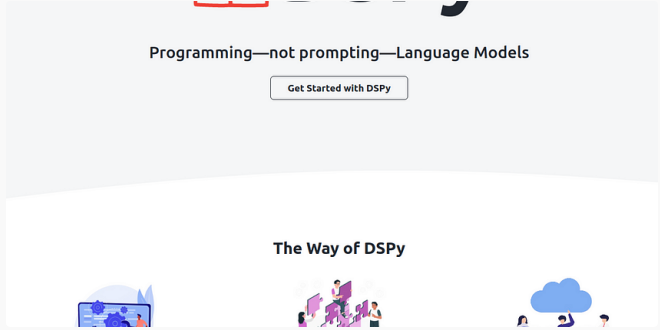
11



1K



14



Vishal Rajput in AIGuys

Prompt Engineering Is Dead: DSPy Is New Paradigm For Prompting

DSPy Paradigm: Let's program—not prompt—LLMs



11 min read · May 29, 2024

See more recommendations