

Open in app ↗



Search

Write



★ Member-only story

Building LLM Applications: Advanced RAG (Part 10)



Vipra Singh · Follow

48 min read · Apr 27, 2024



384



2



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](#)
8. [Evaluation](#)

9. Serving LLMs

10. *Advanced RAG (This Post)*

Table Of Contents

- 1. Naive RAG
- 1.1. Issues with Naive RAG
 - 1.1.1. Indexing
 - 1.1.2. Retrieval
 - 1.1.3. Generation
- 2. Move Towards Advanced RAG
 - 2.1. Pre-Retrieval Optimization
 - 2.2. Retrieval optimization
 - 2.3. Post-retrieval optimization
- 3. Advanced RAG Techniques
- 4. PRE-RETRIEVAL TECHNIQUES
 - 4.1. PDF Parsing
 - 4.1.1. The Challenges of Parsing PDF
 - 4.1.2. How to parse PDF documents
 - 4.1.3. Rule-based methods
 - 4.1.4. Methods based on deep learning models.
 - 4.2. Context Enrichment
 - 4.2.1. Sentence Window Retrieval
 - 4.2.2. Auto-merging Retriever (aka Parent Document Retriever)
 - 4.3. Query Rewriting
 - 4.3.1. Hypothetical Document Embeddings (HyDE).
 - 4.3.2. Rewrite-Retrieve-Read
 - 4.3.3. Step-Back Prompting
 - 4.3.4. Query2doc
 - 4.3.5. ITER-RETGEN

- 4.4. Semantic Chunking
 - 4.4.1. Embedding-based methods
 - 4.4.2. Model-based methods
 - 4.4.2.1. Naive BERT
 - 4.4.2.2. Cross Segment...

vishnu, read this story from Vipra Singh — and all the best stories 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
- ✦ Dive deeper into the topics that matter to you
- ✦ Get in-depth articles answering thousands of questions
- ✦ Achieve your personal and professional goals



Jonathan Lethem
New York Times Best-Selling
Author



Susan Orlean
Staff Writer
The New Yorker



Dr. Tom Frieden
Former CDC Director



Savala Nolan
Professor, UC Berkeley School
of Law



Roger Martin
Professor, Strategy Advisor,
Former Dean



Julie Zhuo
Former VP of Product Design,
Facebook

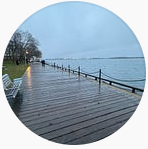


Ryan Holiday
Best-Selling Author and
Entrepreneur



Laura Vanderkam
Best-Selling Author, TED
Speaker

Upgrade



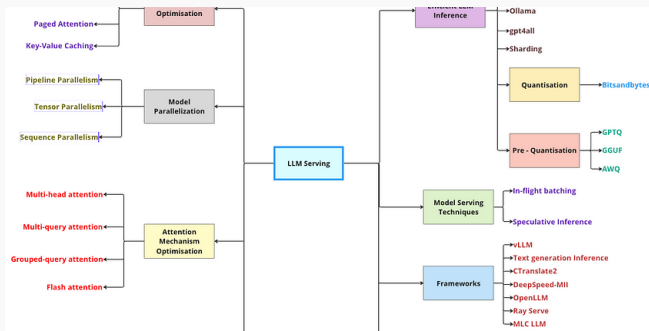
Written by Vipra Singh

1K Followers

Follow



More from Vipra Singh



Vipra Singh

Building LLM Applications: Serving LLMs (Part 9)

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

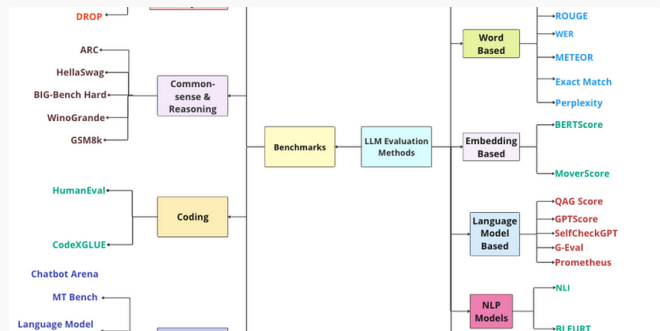
🌟 · 50 min read · Apr 17, 2024



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

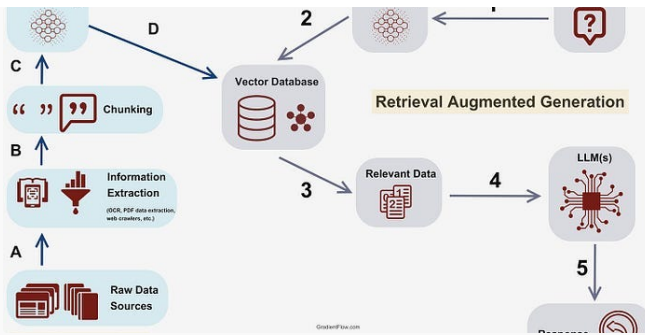


276



1





Vipra Singh

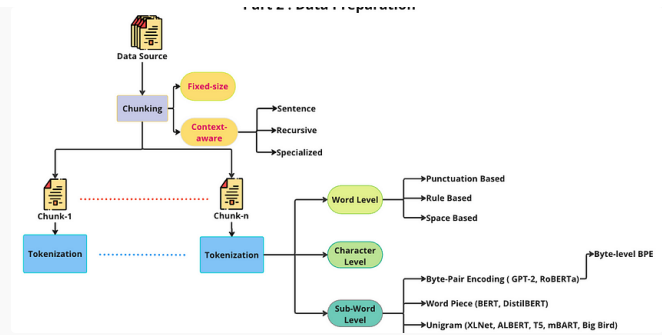
Building LLM Applications: Introduction (Part 1)

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

5 min read · Jan 8, 2024

905 3

...



Vipra Singh

Building LLM Applications: Data Preparation (Part 2)

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

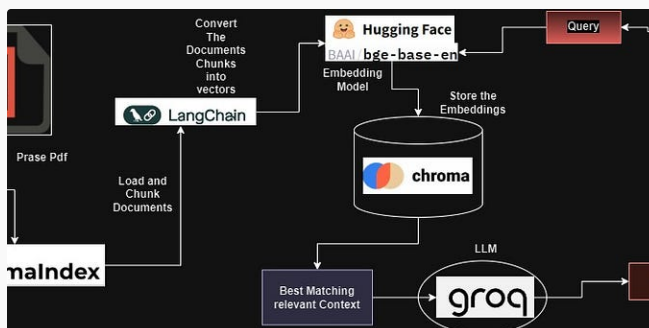
14 min read · Jan 8, 2024

378 2

...

See all from Vipra Singh

Recommended from Medium





Plaban Nayak in The AI Forum

RAG on Complex PDF using LlamaParse, Langchain and Groq

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

13 min read · Apr 7, 2024



678



9



Paul Iusztin in Decoding ML

The 4 Advanced RAG Algorithms You Must Know to Implement

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

15 min read · May 4, 2024



1.4K



10



Lists



Natural Language Processing

1494 stories · 1011 saves



ChatGPT prompts

47 stories · 1642 saves



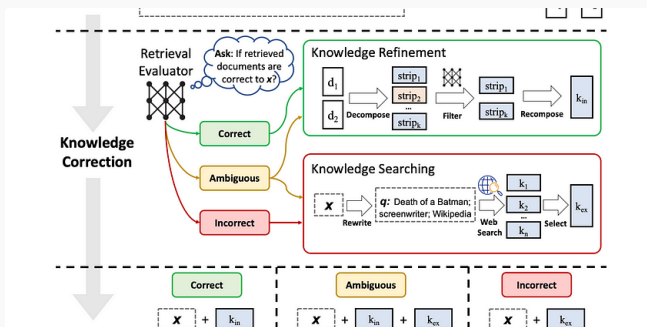
AI Regulation

6 stories · 473 saves



Generative AI Recommended Reading

52 stories · 1106 saves

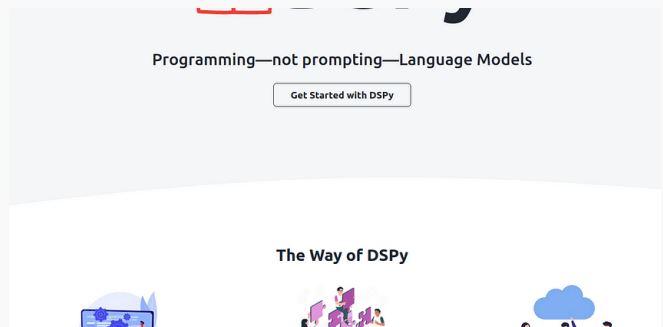


Barsha Rani Swain in GoPenAI

Advanced RAG: Corrective Retrieval Augmented Generation...

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

10 min read · Apr 23, 2024



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



320



1K



14



Ignacio de Gregorio in Towards AI

RAG 2.0, Finally Getting RAG Right!

The Creators of RAG Present its Successor

★ · 9 min read · Apr 10, 2024



2.5K



17



207



2



Ian Kelk

RAG Detective: Retrieval Augmented Generation with...

This article was produced as part of the final project for Harvard's AC215 Fall 2023 course.

16 min read · Dec 10, 2023

See more recommendations