

[Open in app ↗](#)

Search



Write



◆ Member-only story

Building LLM Applications: Evaluation (Part 8)

Vipra Singh · [Follow](#)

48 min read · Apr 7, 2024

276

1



...

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 \(This Post\)](#)

9. Serving LLMs

10. Advanced RAG

Table Of Contents

- 1. Overview
- 2. LLM Benchmarking Vs. Evaluation
- 3. LLM Benchmarking
 - 3.1. Language Understanding and QA Benchmarks
 - 3.1.1. TruthfulQA
 - 3.1.2. MMLU (Massive Multitask Language Understanding)
 - 3.1.3. DROP
 - 3.2. Common-sense and Reasoning Benchmarks
 - 3.2.1. ARC (AI2 Reasoning Challenge)
 - 3.2.2. HellaSwag
 - 3.2.3. BIG-Bench Hard (Beyond the Imitation Game Benchmark)
 - 3.2.4. WinoGrande
 - 3.2.5. GSM8k
 - 3.3. Coding Benchmarks
 - 3.3.1. HumanEval
 - 3.3.2. CodeXGLUE
 - 3.4. Conversation and Chatbot Benchmarks
 - 3.4.1. Chatbot Arena (by LMSys)
 - 3.4.2. MT Bench
 - 3.4.3. Language Model Evaluation Harness (by EleutherAI)
 - 3.4.4. Stanford HELM
 - 3.4.5. PromptBench (by Microsoft)
 - 4. Limitations of LLM Benchmarks
 - 5. LLM Evaluation Metrics
 - 6. Different Ways to Compute Metric Scores

• 6.1. Statistical Scorers

- 6.1.1. Word Error Rate (WER)
- 6.1.2. Exact match
- 6.1.3. Perplexity

◦ ...

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](#)

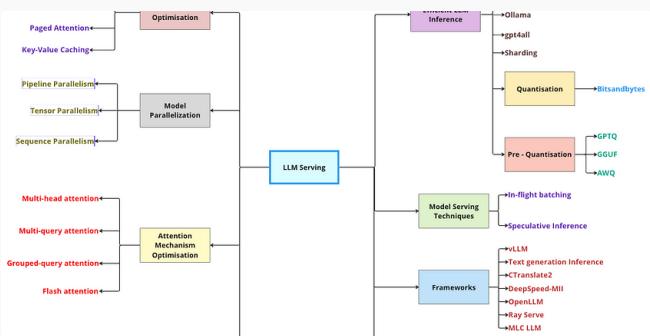
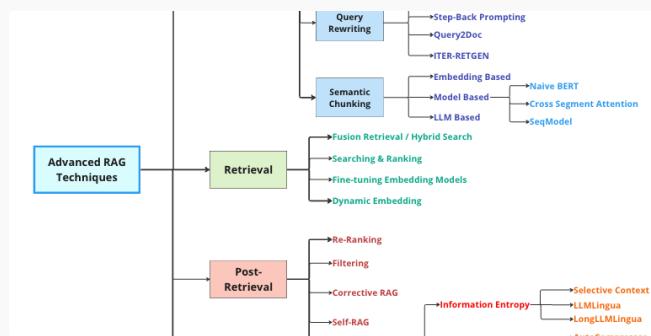


Written by Vipra Singh

1K Followers

[Follow](#)


More from Vipra Singh



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



...



Building LLM Applications: Serving LLMs (Part 9)

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

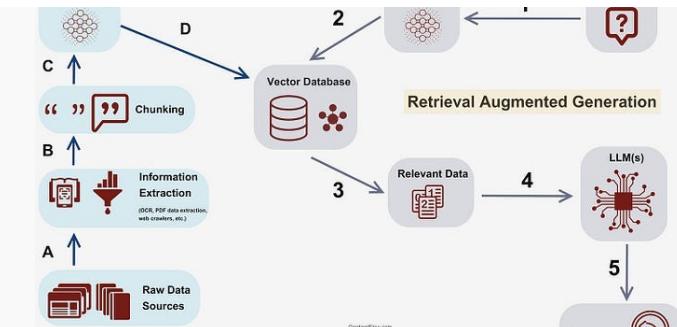
★ · 50 min read · Apr 17, 2024

546

3



...



Vipra Singh

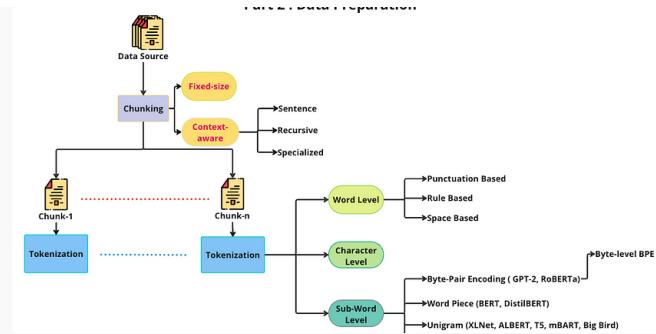
Building LLM Applications: Introduction (Part 1)

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...

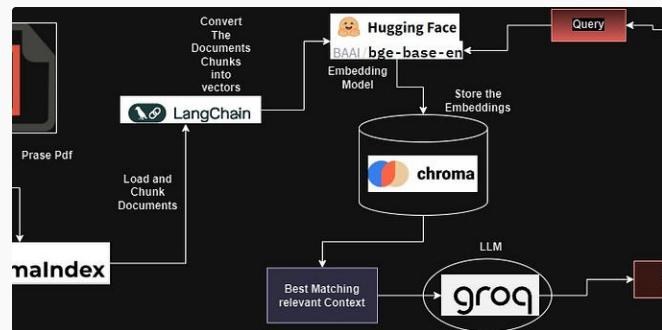
14 min read · Jan 8, 2024

378 2

...

[See all from Vipra Singh](#)

Recommended from Medium





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



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 Languag...

13 min read · Apr 7, 2024

678

9



Lists



Natural Language Processing

1494 stories · 1011 saves



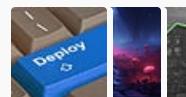
ChatGPT prompts

47 stories · 1642 saves



AI Regulation

6 stories · 473 saves



Predictive Modeling w/ Python

20 stories · 1254 saves

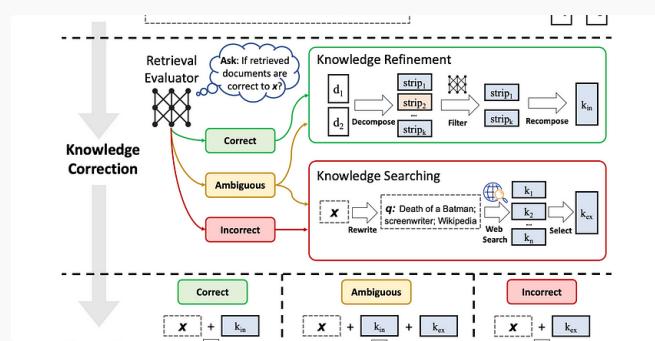


Jane Huang in Data Science at Microsoft

Evaluating LLM systems: Metrics, challenges, and best practices

A detailed consideration of approaches to evaluation and selection

11 min read · Mar 5, 2024



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

1.1K

12

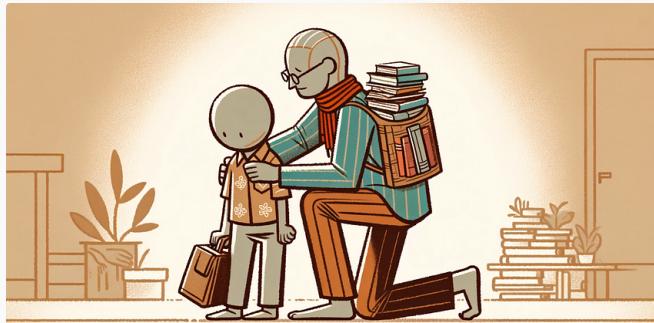


...

320



...



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



...

[See more recommendations](#)