Open in app ↗

Search Write

✨ Member-only story

Building LLM Applications: Serving LLMs (Part 9)

Vipra Singh · Follow

50 min read · Apr 17, 2024

546 3

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 (This Post)*
10. *Advanced RAG*

Table Of Contents

- 1. Run LLMs locally
  - 1.1. Open-source LLMs
- 2. Load LLMs Efficiently
  - 2.1. HuggingFace
  - 2.2. LangChain
  - 2.3. Llama.cpp
  - 2.4. Llamafile
  - 2.5. Ollama
  - 2.6. GPT4ALL
  - 2.7. Sharding
  - 2.8. Quantize with Bitsandbytes
  - 2.9. Pre-Quantization (GPTQ vs. AWQ vs. GGUF)
- 3. Inference Optimization
- 4. Understanding LLM inference
  - 4.1. Prefill phase or processing the input
  - 4.2. Decode phase or generating the output
  - 4.3. Request batching
  - 4.4. Continuous batching
  - 4.5. PagedAttention: A Memory-Centric Solution
  - 4.6. Key-value caching
  - 4.6.1. LLM memory requirement
- 5. Scaling up LLMs with model parallelization
  - 5.1. Pipeline parallelism
  - 5.2. Tensor parallelism
  - 5.3. Sequence parallelism

- 6. Optimizing the attention mechanism
  - 6.1. Multi-head attention
  - 6.2. Multi-query attention
  - 6.3. Grouped-query attention
  - 6.4. Flash attention
  - 6.5. Efficient management...

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

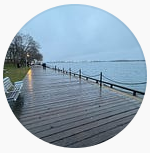
Roger Martin
Professor, Strategy Advisor, Former Dean

Julie Zhuo
Former VP of Product Design, Facebook

Ryan Holiday
Best-Selling Author and Entrepreneur

Laura Vanderkam
Best-Selling Author, TED Speaker

Upgrade

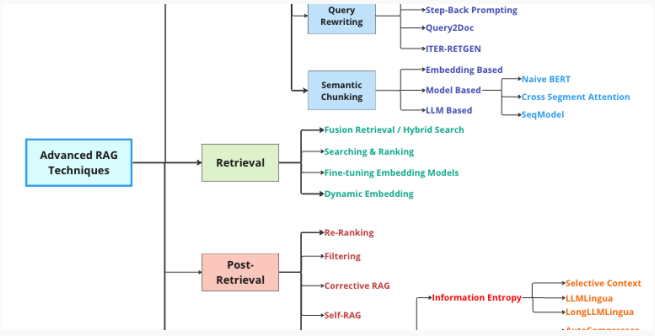


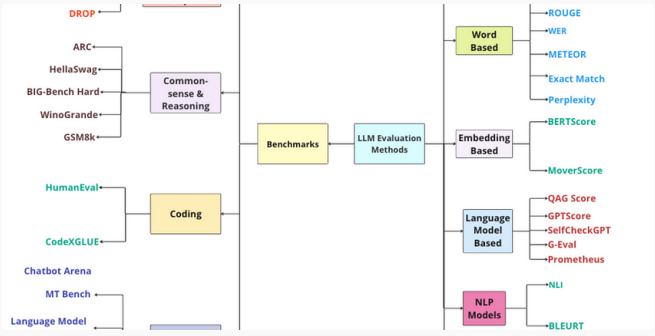
Written by Vipra Singh

1K Followers

Follow

More from Vipra Singh





Vipra Singh

Building LLM Applications: Advanced RAG (Part 10)

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

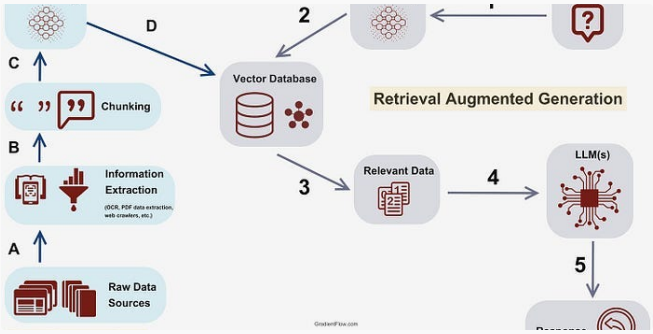
Vipra Singh

Building LLM Applications: Evaluation (Part 8)

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: Introduction (Part 1)

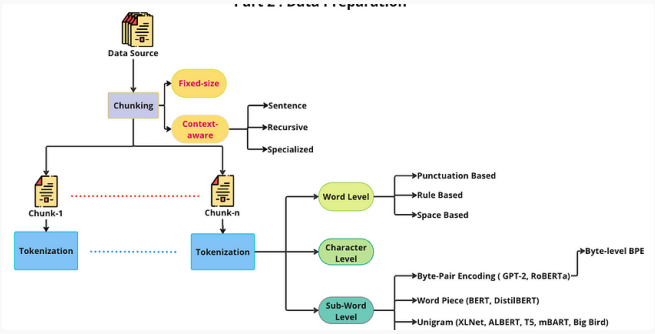
Learn Large Language Models (LLM) through the lens of a Retrieval Augmented...

5 min read · Jan 8, 2024

905 3

✦ · 48 min read · Apr 7, 2024

276 1



Vipra Singh

Building LLM Applications: Data Preparation (Part 2)

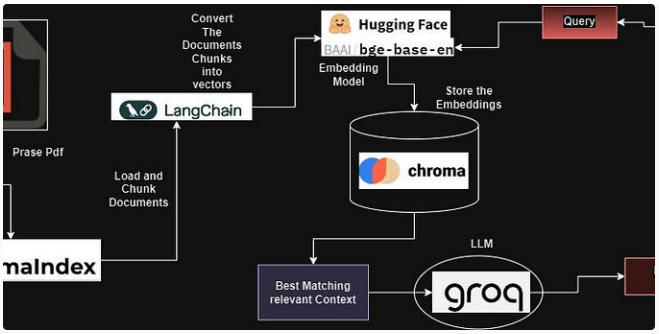
Learn Large Language Models (LLM) through the lens of a Retrieval Augmented...

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 Languag...

13 min read · Apr 7, 2024

678 9

Paul Iusztin in Decoding ML

An End-to-End Framework for Production-Ready LLM Systems b...

From data gathering to productionizing LLMs using LLMOps good practices.

16 min read · Mar 16, 2024

1.7K 11

Lists

Natural Language Processing

1494 stories · 1011 saves

AI Regulation

6 stories · 473 saves

ChatGPT prompts

47 stories · 1642 saves

Generative AI Recommended Reading

52 stories · 1106 saves

Mahesh

How to Productionize Large Language Models (LLMs)

Understand LLMOps, architectural patterns, how to evaluate, fine tune & deploy...

94 min read · Mar 27, 2024

186 2

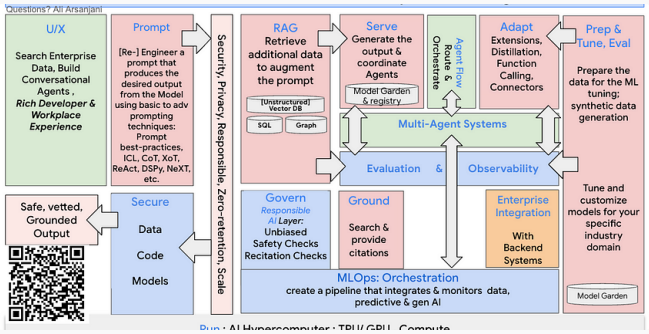
Ignacio de Gregorio in Towards AI

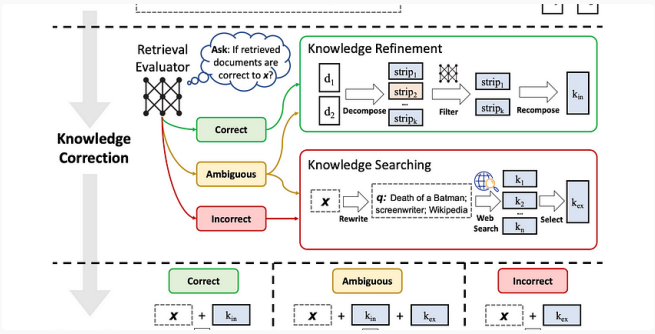
RAG 2.0, Finally Getting RAG Right!

The Creators of RAG Present its Successor

✦ · 9 min read · Apr 10, 2024

2.5K 17





Ali Arsanjani

The GenAI Reference Architecture

26 min read · Apr 28, 2024

941 6

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

320

See more recommendations