University of Applied Sciences and Arts Northwestern Switzerland
School of Engineering

member of swissuniversities

RAG
Retrieval Augmented Generation

NPR Mini-Challenge 1 - Introduction

27. February 2025

George Rowlands & Daniel Perruchoud
Institute for Data Science (I4DS)

Image generated by DALL-E 3

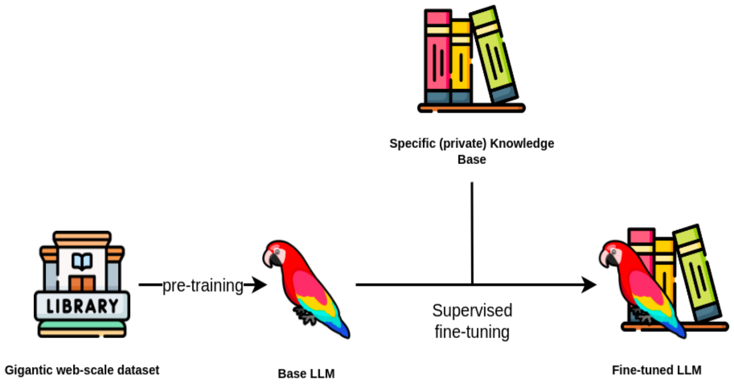
Content

- **Motivation**
  - Fine-tuning vs RAG
- **Implementation & Challenges**
  - Ingestion
  - Retrieval
  - Generation
- **Evaluation**
  - Using LLMs to test LLMs
- **Advanced Methods**
  - Conversational Memory
  - Chunking & Retrieval
- **Tools & Resources**

Motivation

- We use Large Language Models (LLMs) with billions of parameters trained on a general corpus for weeks every day
- Limitations
  - Inefficient when dealing with specific knowledge
  - Missing trace to knowledge source
  - Using potentially outdated information
- What if we want to add our own curated "domain knowledge" to an LLM?

→ **Possible Solution: Fine-tuning an LLM**



Supervised Fine-tuning: customizing LLMs

Cost of Pre-training for LLama 2

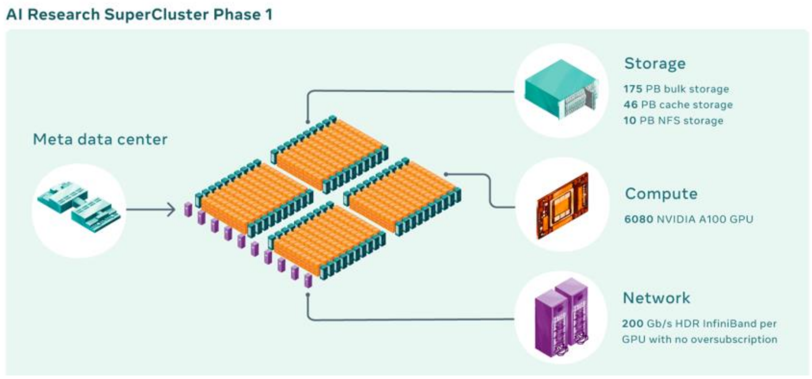
What do you estimate is the cost of pre-training LLama 2 with 7B parameters?

		Time (GPU hours)	Power Consumption (W)	Carbon Emitted (tCO_2eq)
LLAMA 2	7B	184320	400	31.22
	13B	368640	400	62.44
	34B	1038336	350	153.90
	70B	1720320	400	291.42
Total		3311616		539.00

Llama 2: Open Foundation and Fine-Tuned Chat Models

Cost of Pre-training for LLama 2

- 184’320 h × 400 W = 73’728’000 Wh
- Brugg IBB Cost: 0.2947 CHF/kWh
- 73’728 kWh × 0.2947 CHF/kWh = **21’704.14 CHF to pre-train 7B LLama 2.**
- **Only electricity cost!** Hardware & maintenance missing, used Meta RSC + Internal Cluster.

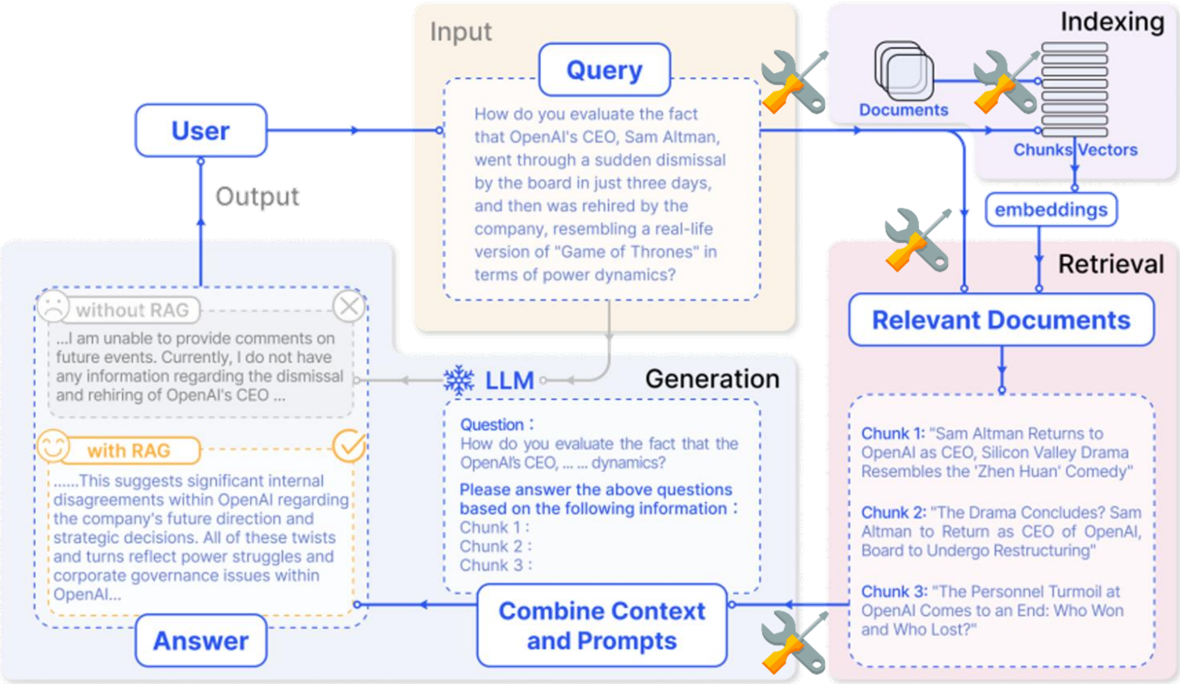


Introducing Meta’s RSC

		Time (GPU hours)	Power Consumption (W)	Carbon Emitted (tCO_2eq)
LLAMA 2	7B	184320	400	31.22
	13B	368640	400	62.44
	34B	1038336	350	153.90
	70B	1720320	400	291.42
Total		3311616		539.00

Llama 2: Open Foundation and Fine-Tuned Chat Models

Retrieval Augmented Generation (RAG) to the rescue



[Retrieval-Augmented Generation for Large Language Models: A Survey]

Retrieval Augmented Generation (RAG) to the rescue

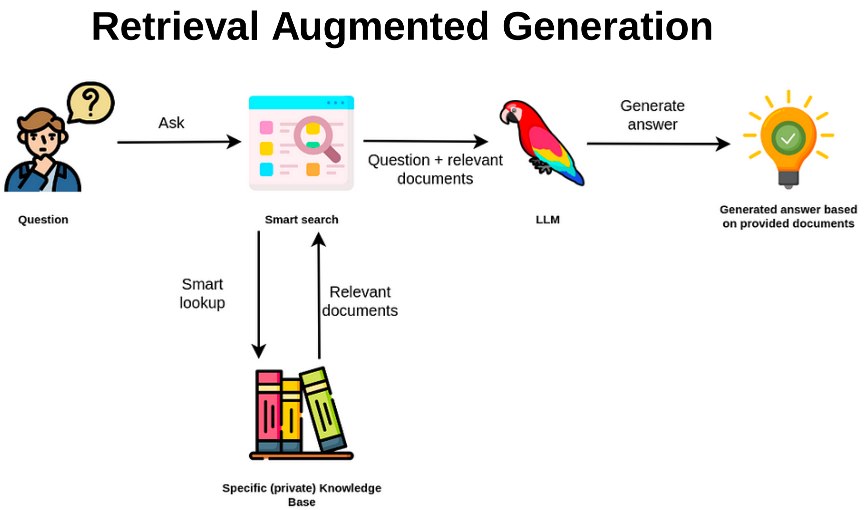
- **Indexing: Create the knowledge base**
  - Documents can be HTML, PDF, etc.
  - Store Embedded chunks are in VectorDB.

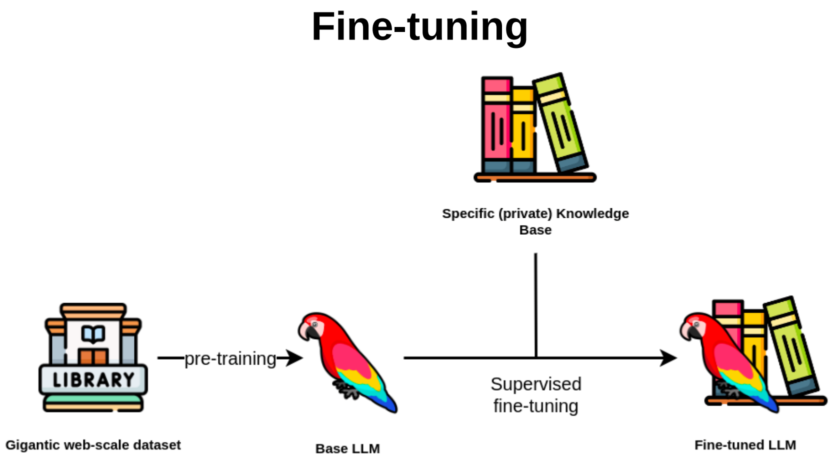
- **Retrieval: Retrieve relevant chunks**
  - Query is embedded and NNs are retrieved from VectorDB.

- **Generation: Synthesize answer**
  - Relevant Chunks are given to the LLM to augment the answer generation "**Context aware prompt**".

Retrieval-Augmented Generation for Large Language Models: A Survey

Retrieval Augmented Generation (RAG) to the rescue



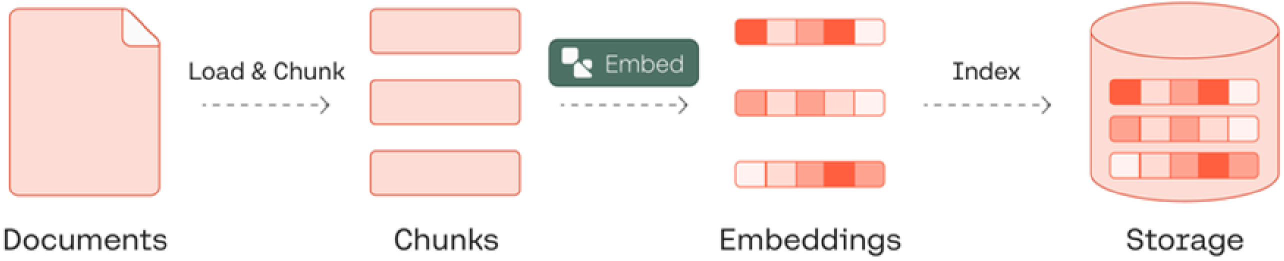


- RAG doesn’t require any training, answers “fact based” with domain-specific references, protects data privacy, requires prompt engineering.

Implementation & Challenges

Ingestion

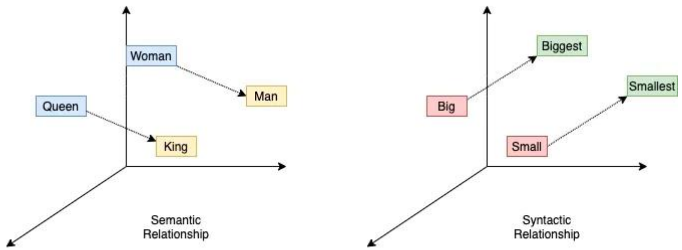
- **Pre-processing**: Data needs to be collected and processed.
  - Challenges: Scanned documents, formulas, tables, images, duplicates etc.
- **Enrichment**: Add metadata for filtering, file name, language, date of publication etc.
- **Chunking**: Documents need to be chunked for easy retrieval, whilst keeping data locality.
  - Challenges: Where to chunk, how large should chunks be, tokens vs characters?



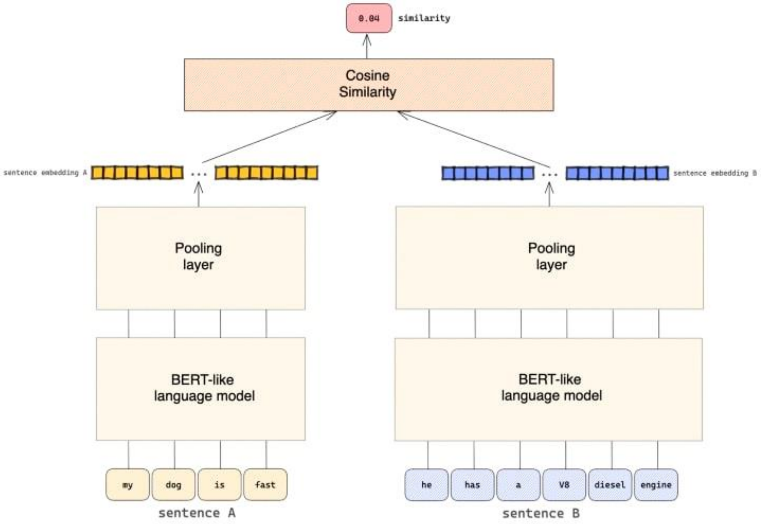
[How to Build a RAG-Powered Chatbot with Chat, Embed, and Rerank]

Ingestion

- **Embedding**: Text that often appears together should be close in the vector space. Now on "sentence" level not on "words".
- Challenges: Context size & chunk size.



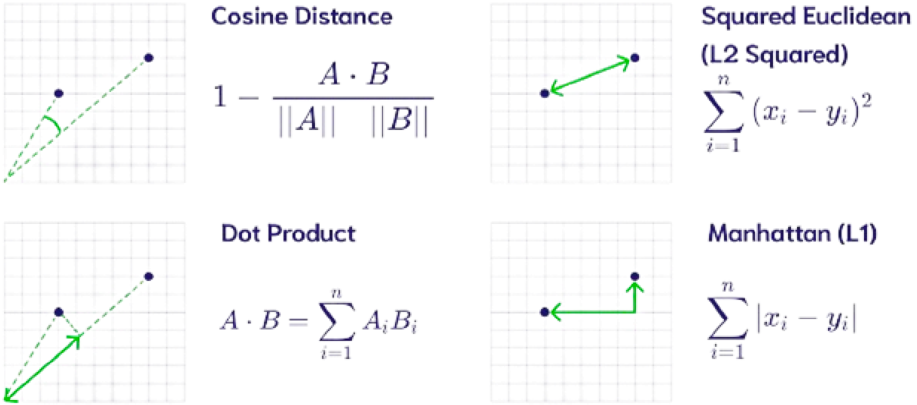
Word2Vec Explained



Large Language Models: SBERT

Retrieval

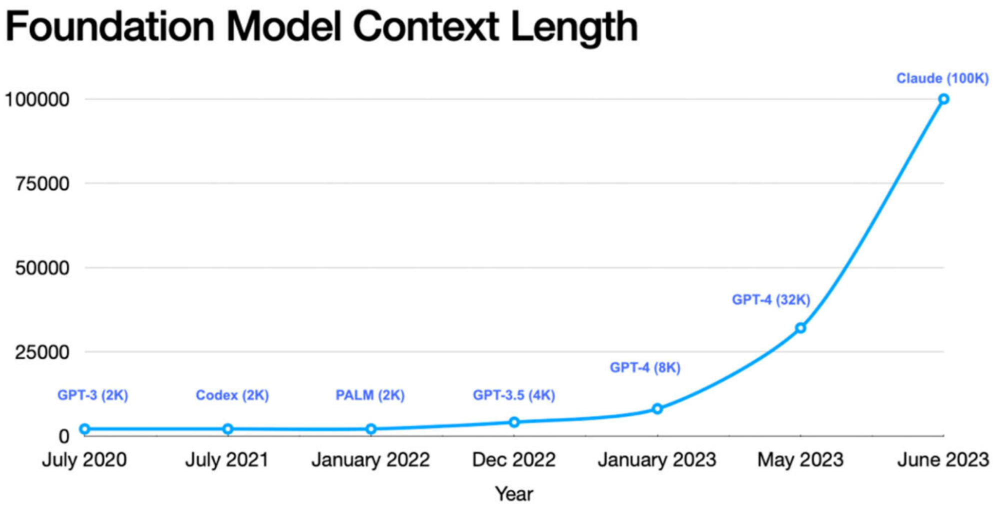
- **Retrieval:** Retrieve the relevant chunks to the query.
- Challenges:
  - How do we know what is relevant? (Choice of distance metric)
  - How much do we retrieve? (Top-k vs confidence threshold)



Distance Metrics in Vector Search

Generation

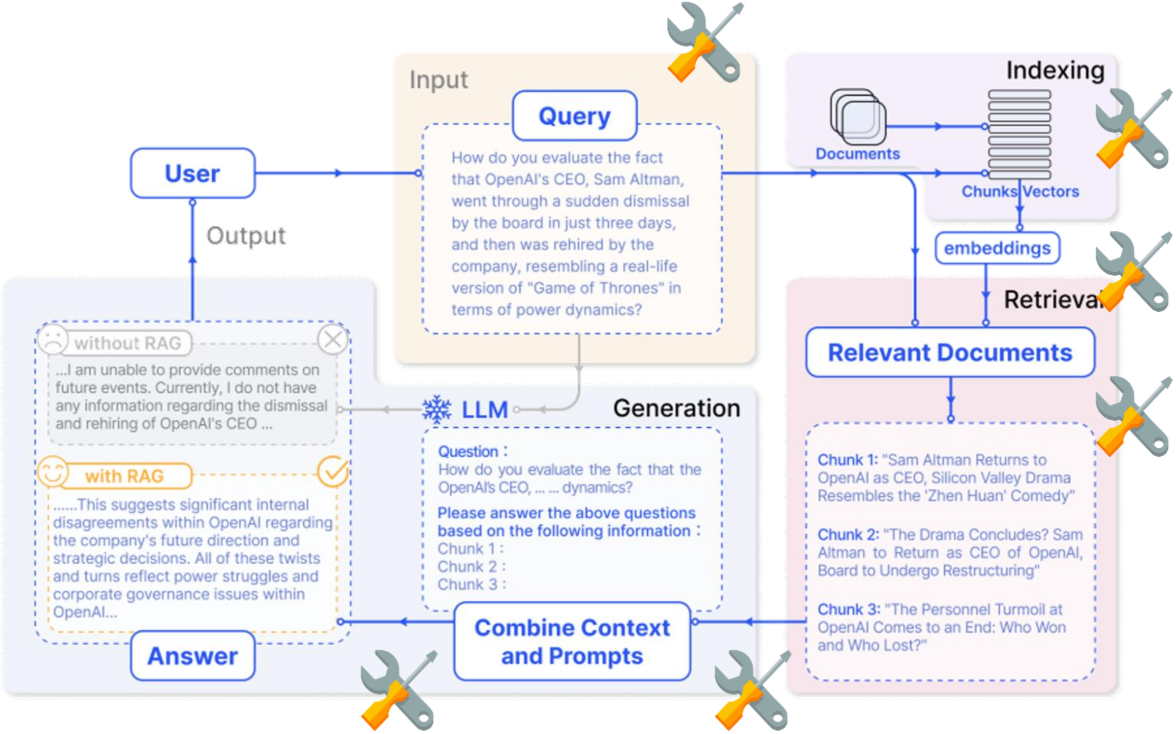
- **Generation:** Synthesize the answer from the given context.
- Challenges:
  - How to reconcile chunk size and LLM token limit?
  - Which prompts to use?



[Variable Sequence Length Training for Long-Context LLMs]

Evaluation

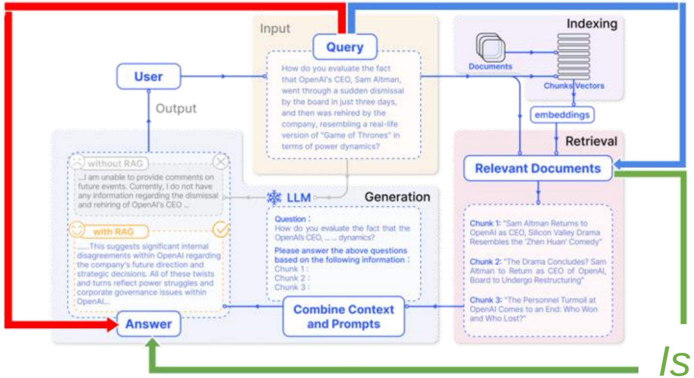
Evaluation – Find optimal RAG configuration



Evaluation

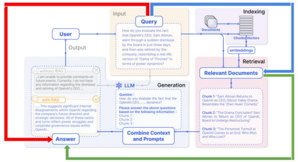
Is the retrieved context relevant to the query?

Is the answer relevant to the query?



Is the answer supported by the retrieved context?

- Use human gold-standard query-answer pairs
- Use LLM to evaluate optimal configuration with synthetic queries



Evaluation

- Context relevance - *Is the retrieved context relevant to the query?*

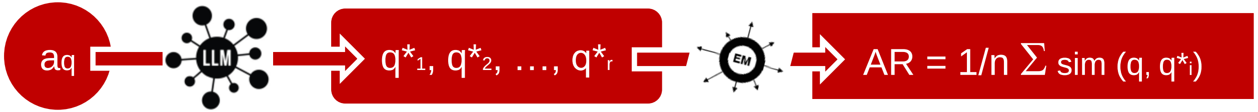
Prompt: Please extract relevant sentences from the provided context that can potentially help answer the following query. …

- Faithfulness - *Is the answer supported by the retrieved context?*

Prompt: …create ≥ 1 statements form each sentence in the answer …

Prompt: Consider statements and determine whether they are supported by the information given in the context …

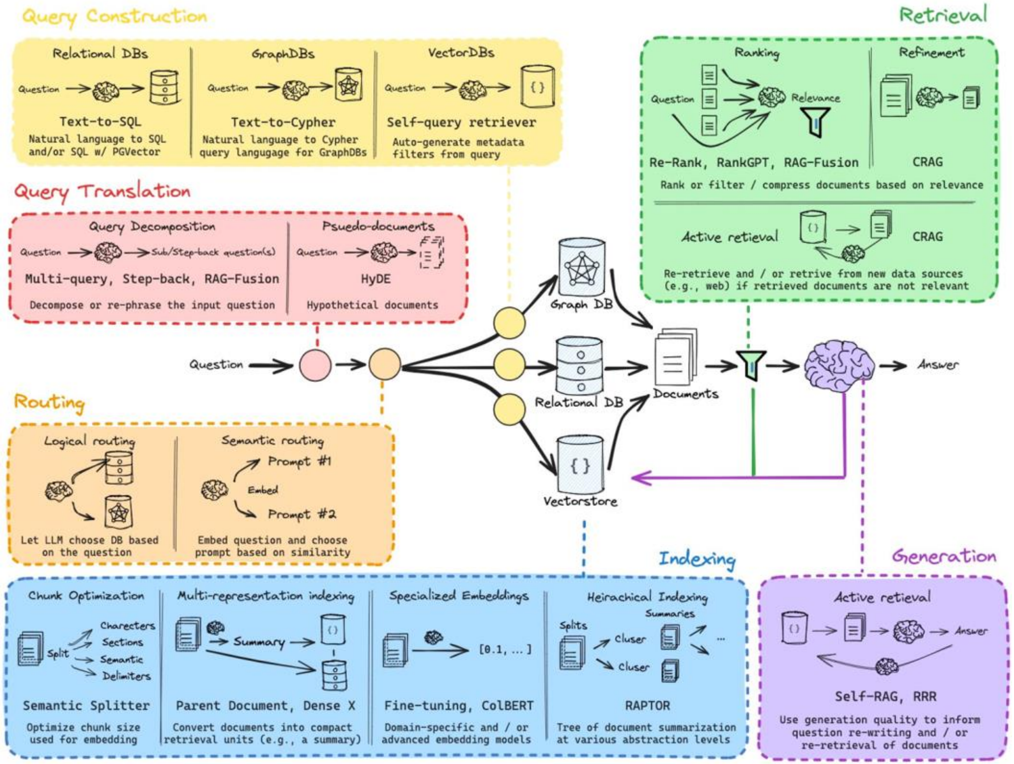
- Answer relevance - *Is the answer relevant to the query?*



Prompt: Generate a query for the given answer …

Advanced Methods

Building Advanced RAG Systems



Tools & Resources

Langchain vs LlamaIndex

LangChain

- Simple chaining of components
- More programming
- More customization and control
- Large community
- Not just RAG

LlamaIndex

- A lot of algorithms already implemented
- Good default implementations
- Quick and less code
- Sometimes a black box
- Harder to customize if even possible

Resources

- Pinecone Library
- LangChain Blog & Docs
- DeepLearning.ai Courses
  - LangChain for LLM Application Development
  - LangChain Chat with Your Data
  - Functions, Tools and Agents with LangChain
  - Advanced Retrieval for AI with Chroma
  - Building and Evaluating Advanced RAG Applications