Who Am I?

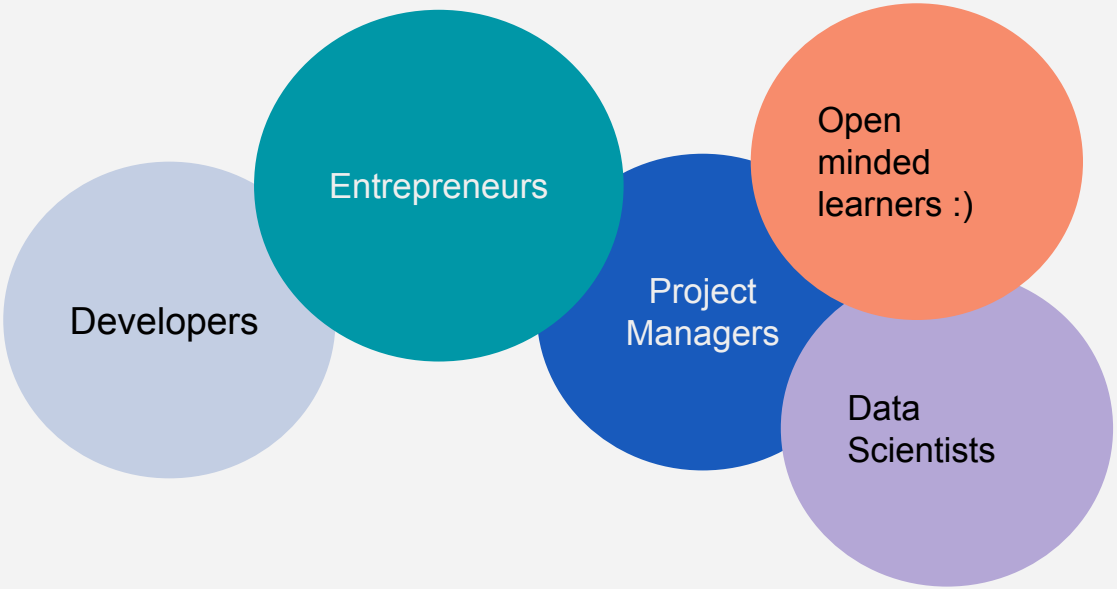
Paulo Dichone

Software, Cloud, AI Engineer
and Instructor

What Is This Course About?

- RAG - Intermediate to Advanced Concepts and Techniques
  - Hands-on on advanced RAG
  - Improving RAG Pipelines performance using advanced RAG techniques

Who Is This Course For



Course Prerequisites

1. Know Programming (highly *preferred…)*
    a. *There will be some Python code*
2. *This is not a programming course*
3. *This is not AI/RAG 101 course*
    a. *Must have some fundamentals on RAG, LLM, Vector databases, etc*
4. Willingness to learn :)

Course Structure

Development Environment setup

- Python
- VS Code (or any other code editor)
- OpenAI API Account and API Key

Set up OpenAI API Account

OpenAI API - Dev Environment Setup

Python (Win, Mac, Linux)

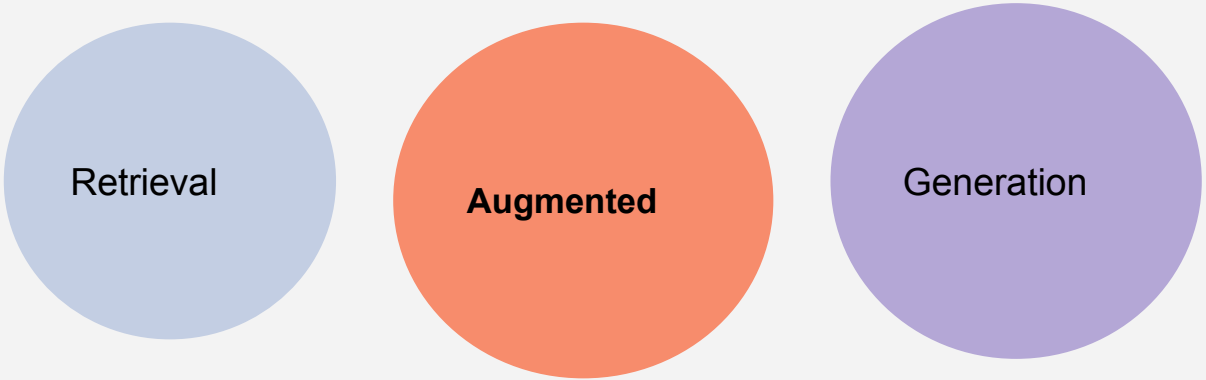
https://kinsta.com/knowledgebase/install-python/

RAG - Deep Dive

- What is it?
- Why (motivation)?
- Advantages

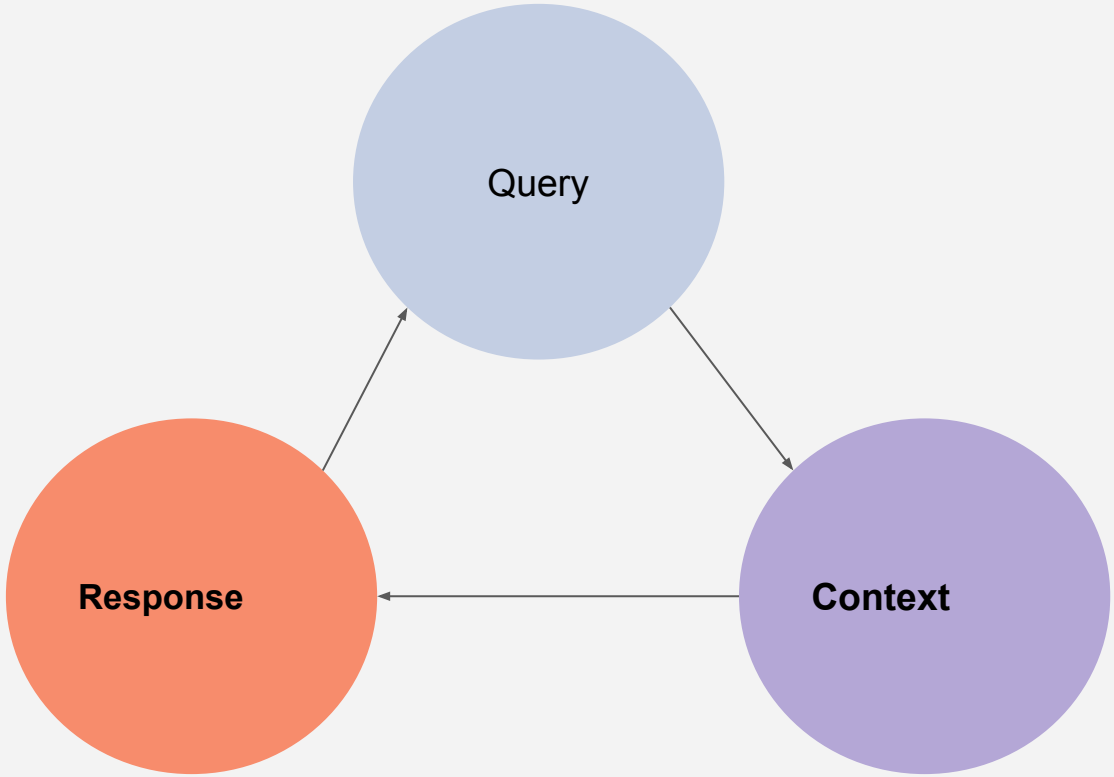
What is a RAG?

Retrieval-Augmented Generation



RAG Triad

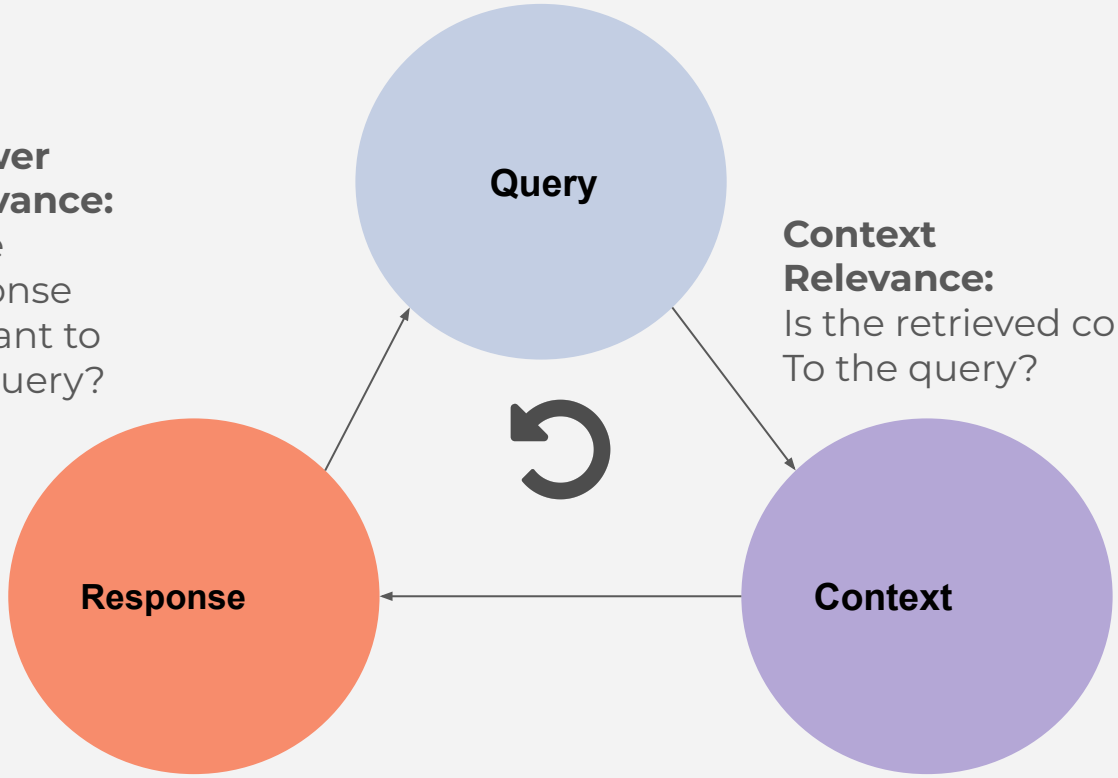
Retrieval-Augmented Generation



The RAG Triad

- Context Relevance
- Groundness
- Answer Relevance

The RAG Triad



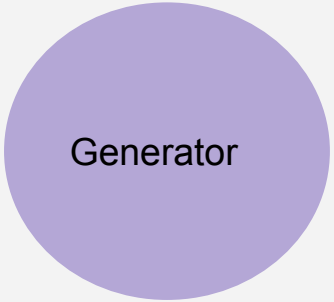
Answer Relevance:
Is the response relevant to the query?

Context Relevance:
Is the retrieved context relevant To the query?

Groundness:
Is the response supported by the context?

Components of RAG

Identifies and retrieves relevant documents

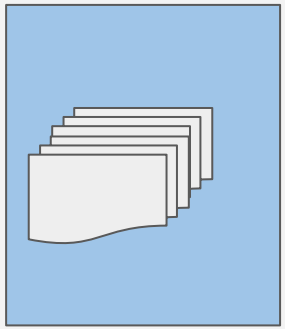


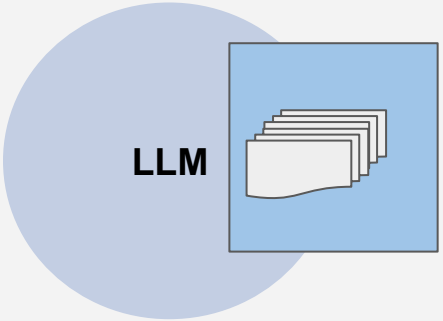
Takes retrieved docs and the input query to generate coherent and contextually relevant response.

What is a RAG?

Definition - *a framework that combines the strengths of retrieval-based systems and generation-based models to produce more accurate and contextual relevant response*

Translation - *... efficient way to customize an LLM (model) with your own data.*

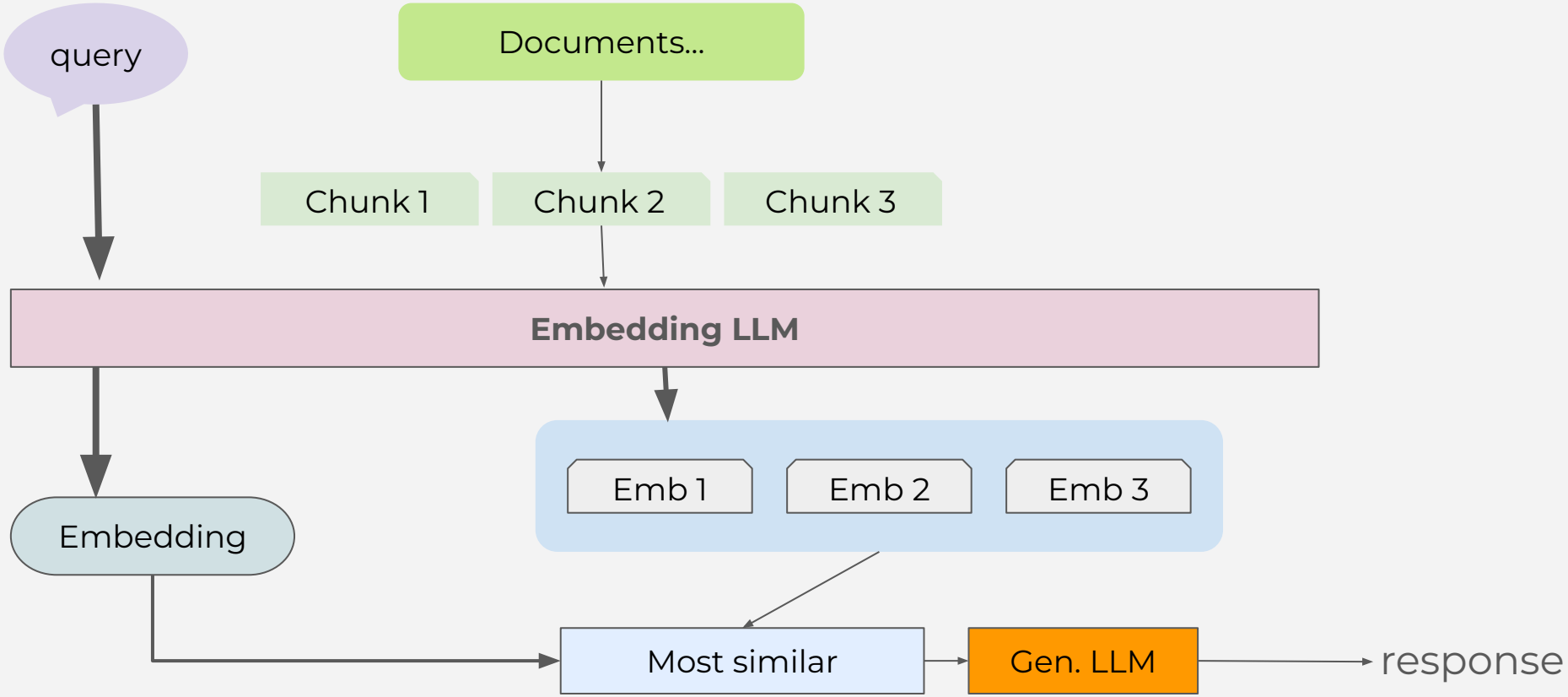




I only know what I was trained on..

Now I also know about specific contextual data :)

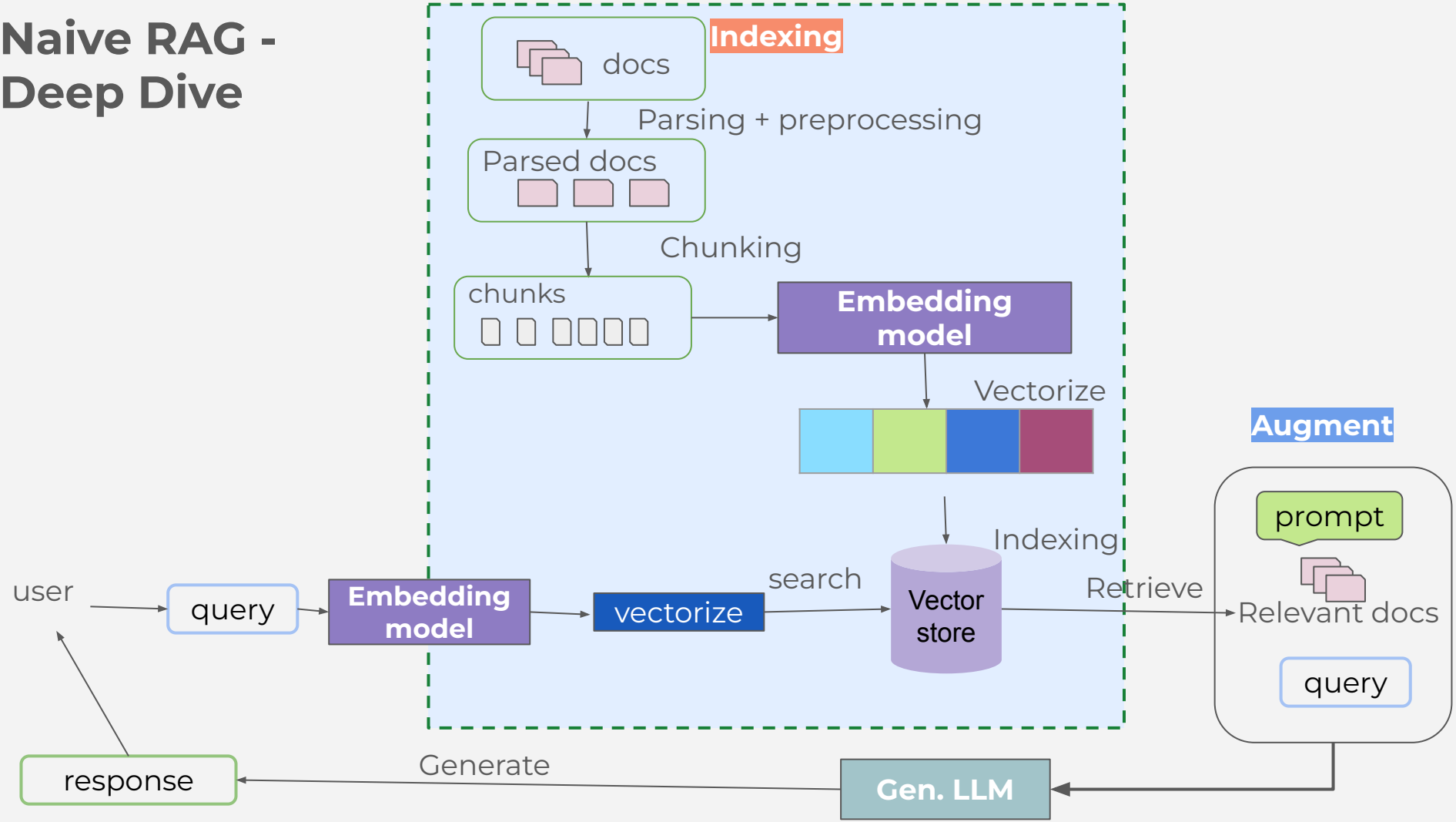
RAG Overview



Naive RAG

Indexing	Retrieval	Generation
Cleaning, extracting data from docs...	Turn question into a vector. Vector comparison... Retrieves closely Related chunks...	The query, choose docs combined into a prompt. The model generates an answer...

Naive RAG - Deep Dive



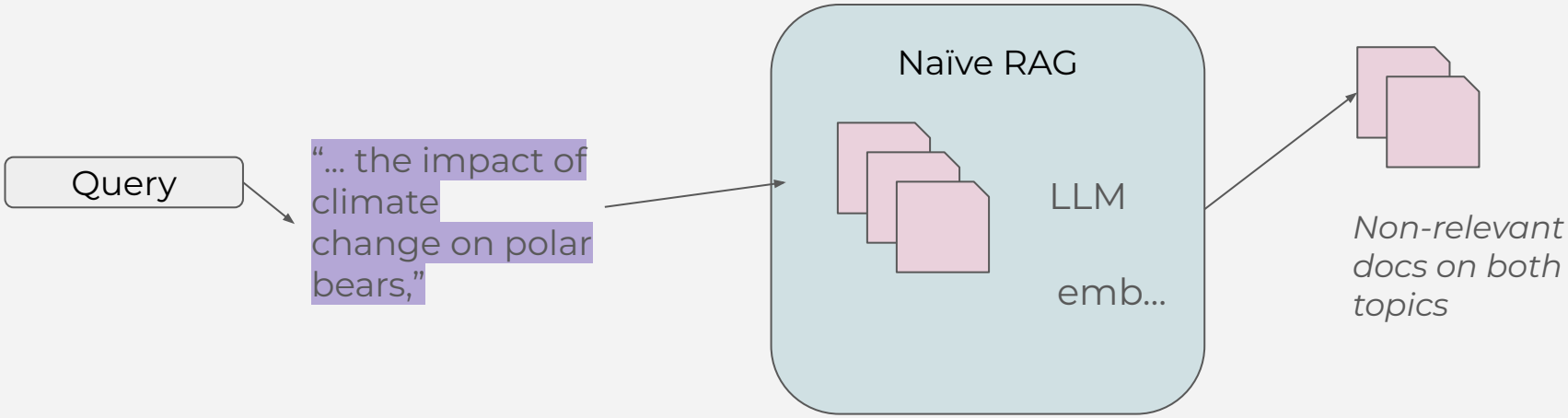
Naive RAG Drawbacks/Challenges/Pitfalls

1. Limited contextual understanding
2. Inconsistent relevance and quality of retrieved documents
3. Poor integration between retrieval and generation
4. Inefficient handling of Large-Scale data
5. Lack of robustness and adaptability

Naive RAG Drawbacks/Challenges/Pitfalls

Limited contextual understanding

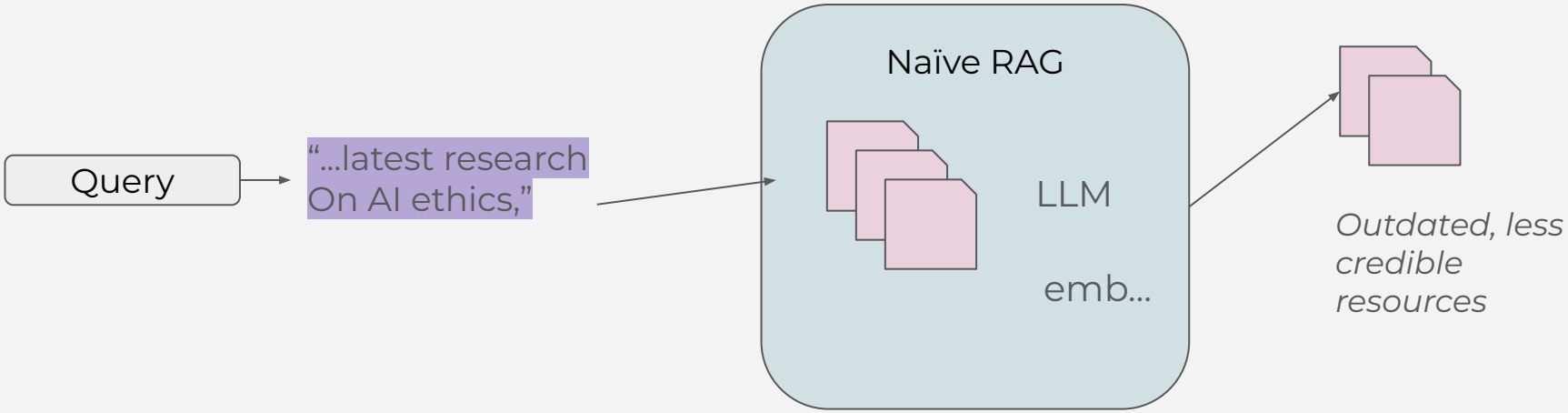
a. Focus on keyword matching or basic semantic search (*retrieving irrelevant or partially relevant documents)*



Naive RAG Drawbacks/Challenges/Pitfalls

Inconsistent relevance and quality of retrieved documents

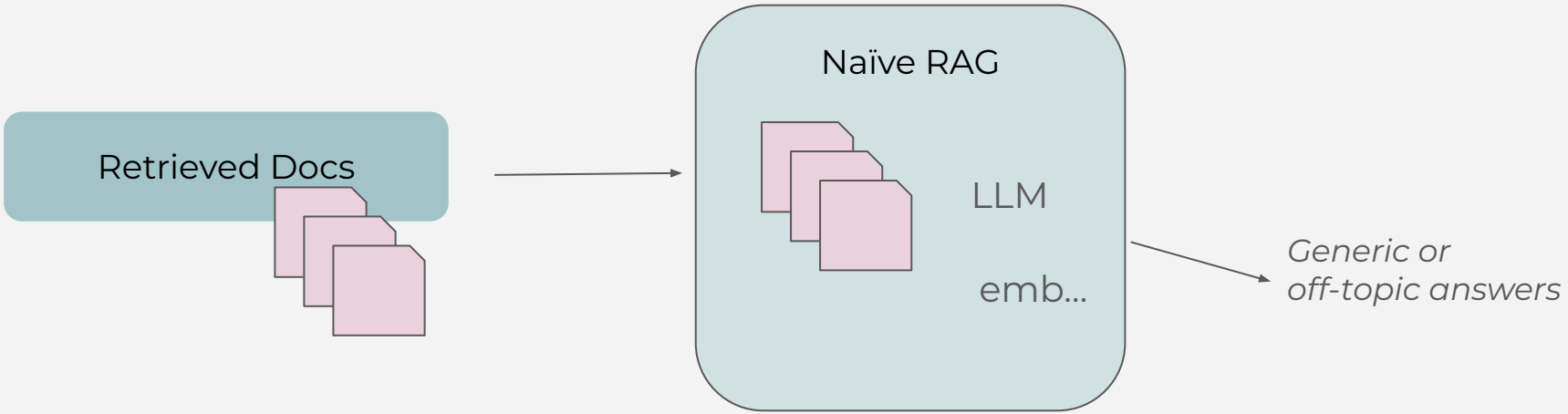
a. Varying in quality and relevance documents (*poor-quality inputs for the gen model)*



Naive RAG Drawbacks/Challenges/Pitfalls

Poor integration between retrieval and generation

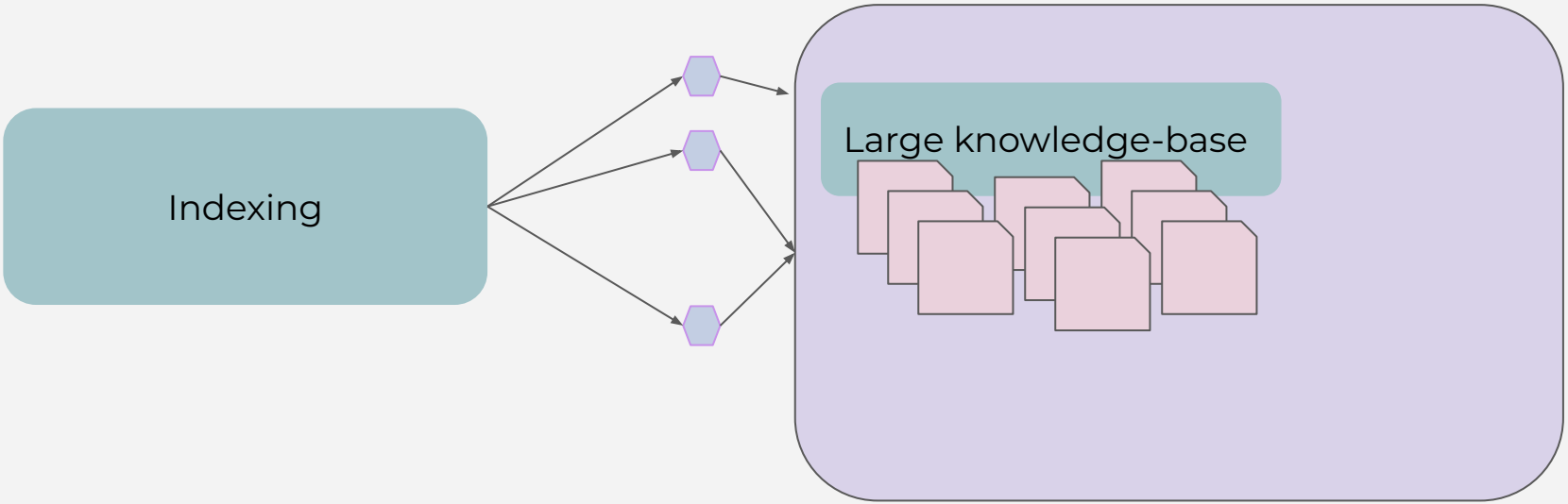
a. Retriever and generator to working in sync (unoptimized information)



Naive RAG Drawbacks/Challenges/Pitfalls

Inefficient handling of Large-Scale data

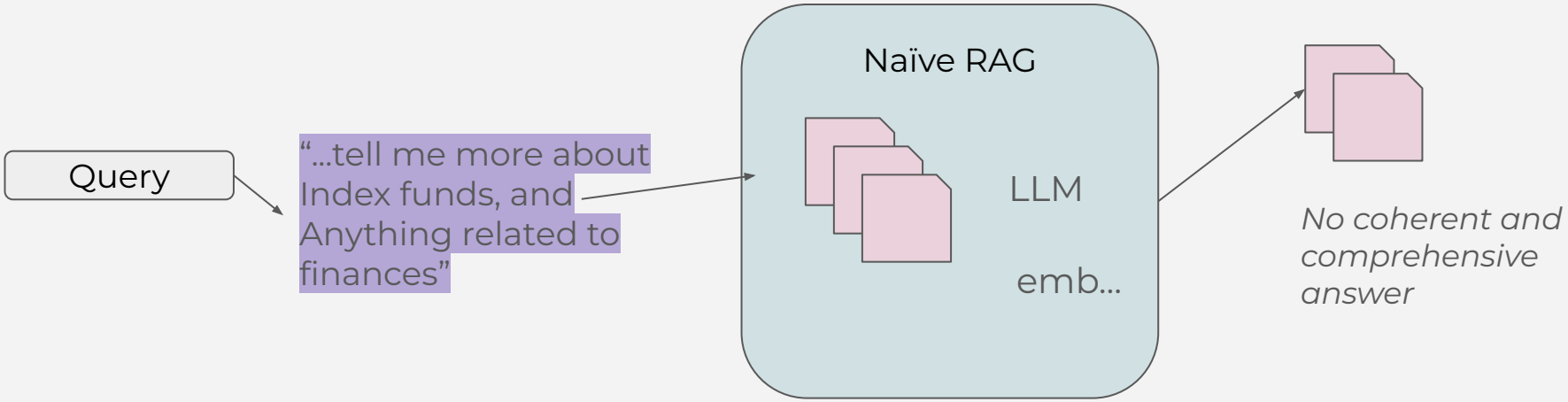
a. Scaling issues, take too long to find relevant docs, or miss critical info due to bad indexing...



Naive RAG Drawbacks/Challenges/Pitfalls

Lack of Robustness and Adaptability

a. Not adaptable to changing contexts or user needs without significant manual intervention



Naive RAG Drawbacks/Challenges/Pitfalls

In summary

We have:

1. **Retrieval Challenges**
    a. Lead to the selection of misaligned or irrelevant chunks, therefore missing of crucial information
2. **Generative Challenges**
    a. The model might struggle with hallucination and have issues with relevance, toxicity or bias in its outputs

Advanced RAG Techniques

And their solutions

Advanced RAG Benefits

Advanced RAG - introduces specific improvements to overcome the limitations of Naive RAG. Focus on enhancing retrieval quality.

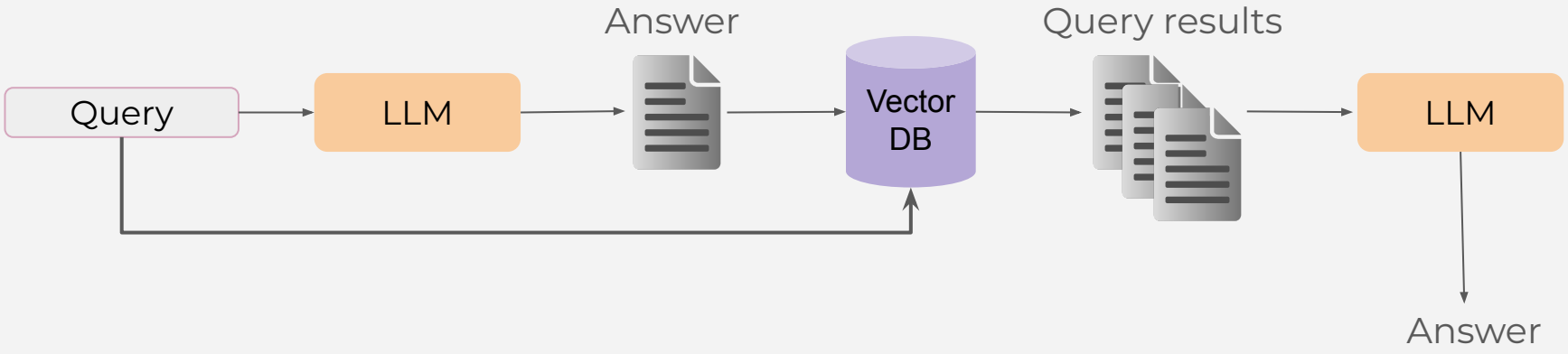
Advanced RAG employs the following strategies:

1. **Pre-retrieval**
    a. Improvement of the indexing structure and user's query
    b. Improves data details, organizing indexes better, adding extra information, aligning things correctly...
2. **Post-retrieval**
    a. Combine pre-retrieval data with the original query
        i. Re-ranking to highlight the most important content...

Advanced RAG Techniques

Query Expansion (*with generated answers)*

Generate potential answers to the query
[*using an LLM]* and to get relevant context



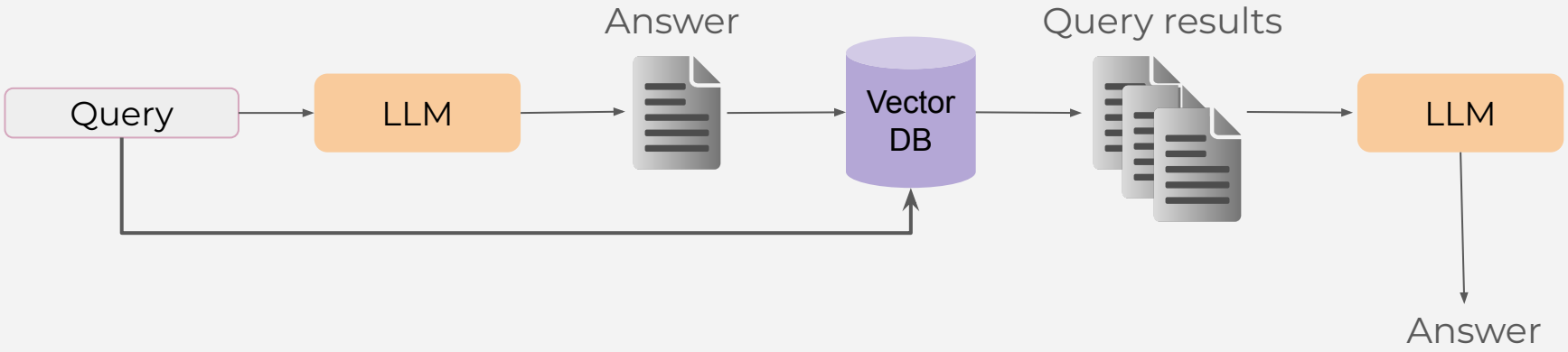
Advanced RAG Techniques

Query Expansion (*with generated answers)*

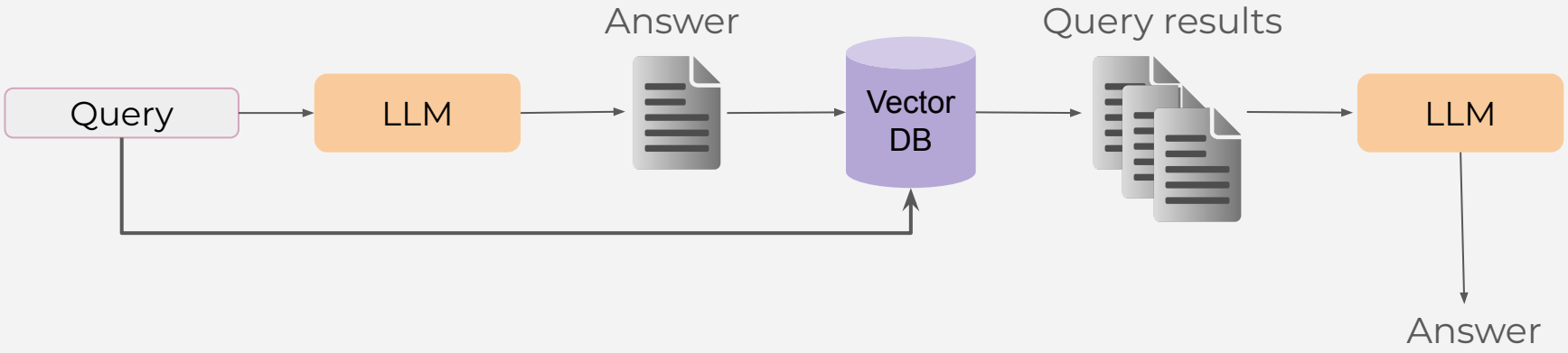
Use cases:

- Information Retrieval
- Question Answering Systems
- E-commerce Search
- Academic Research

Hands-on - Query Expansion



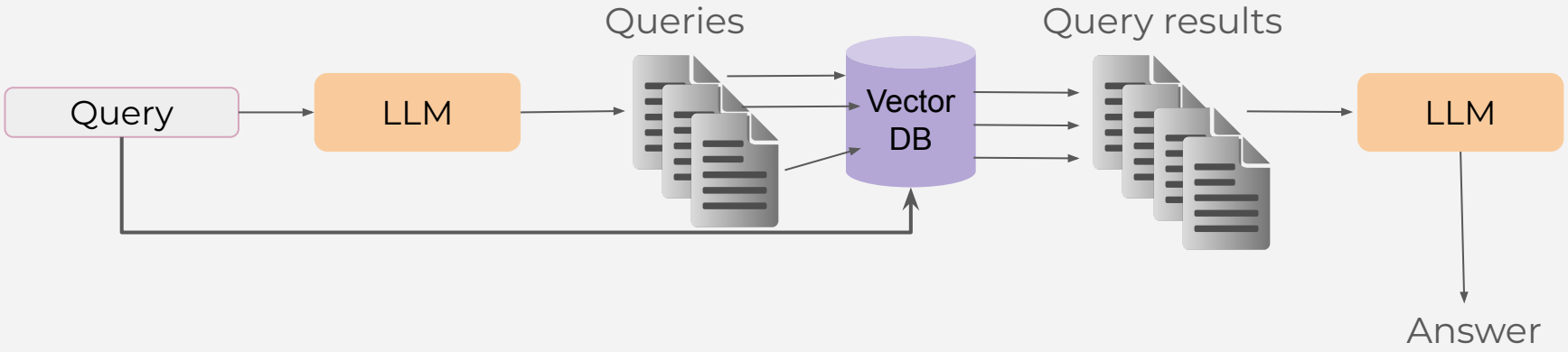
STOP here



Advanced RAG Techniques

Query Expansion (*with multiple queries)*

Use the LLM to generate additional queries that might help getting the most relevant answer.



Advanced RAG Techniques

Query Expansion (*with multiple queries)*

Use cases:

- Exploring Data Analysis
- Academic Research
- Customer Support
- Healthcare Information Systems

Advanced RAG Techniques

Query Expansion (with multiple queries)

Advanced RAG Techniques

Downsides:

- Lots of results
  - queries might not always be relevant or useful
- Results not always relevant and or useful

Hence, we need another technique to find relevant results...

Relevance Feedback or Re-ranking

Advanced RAG Techniques - Expansion with Multi queries

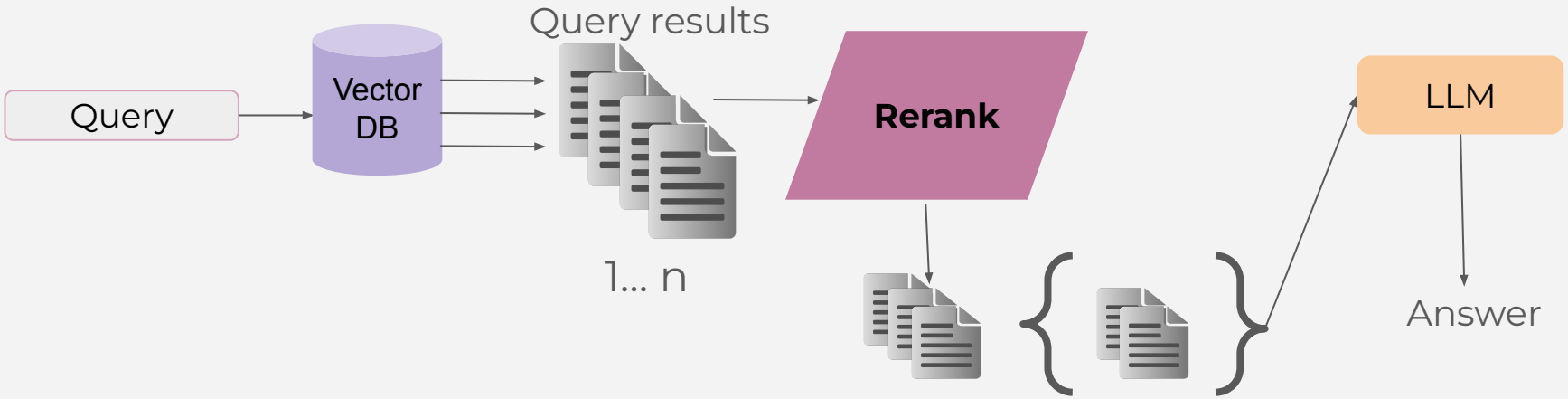
Your turn:

- Play with different prompts and queries
  - See what results you get each time
  - Keep refining the prompt and see the results

Advanced RAG Techniques

Re-ranking(*Cross-encoder)*

Applying a sophisticated process to re-evaluate and reorder the initial retrieved documents



Advanced RAG Techniques

Re-Ranking

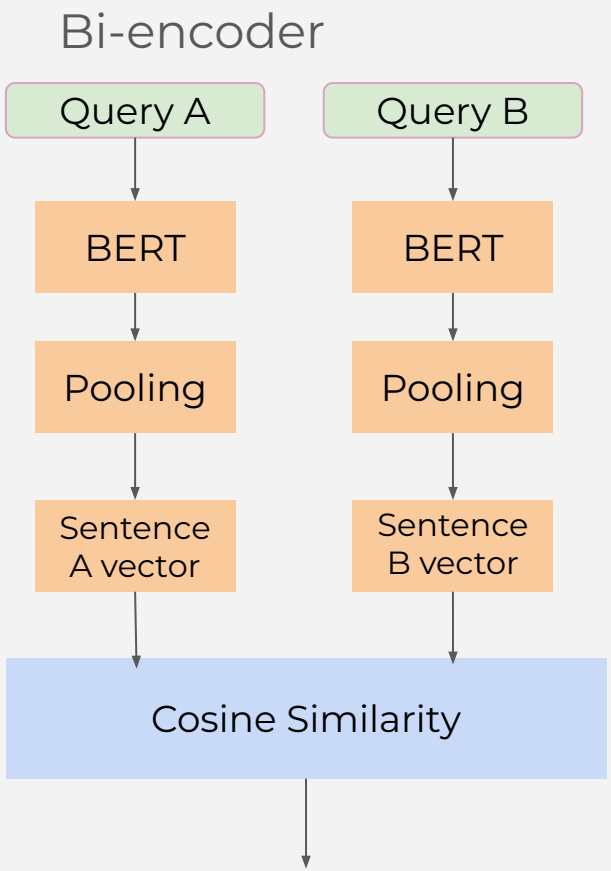
Use cases:

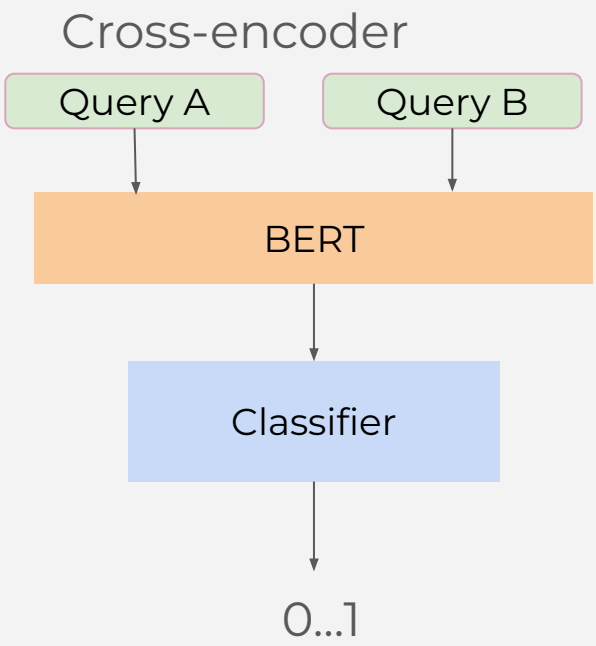
- Search engines
- Question Answering systems
- Recommendation systems
- Legal document search

Hands on

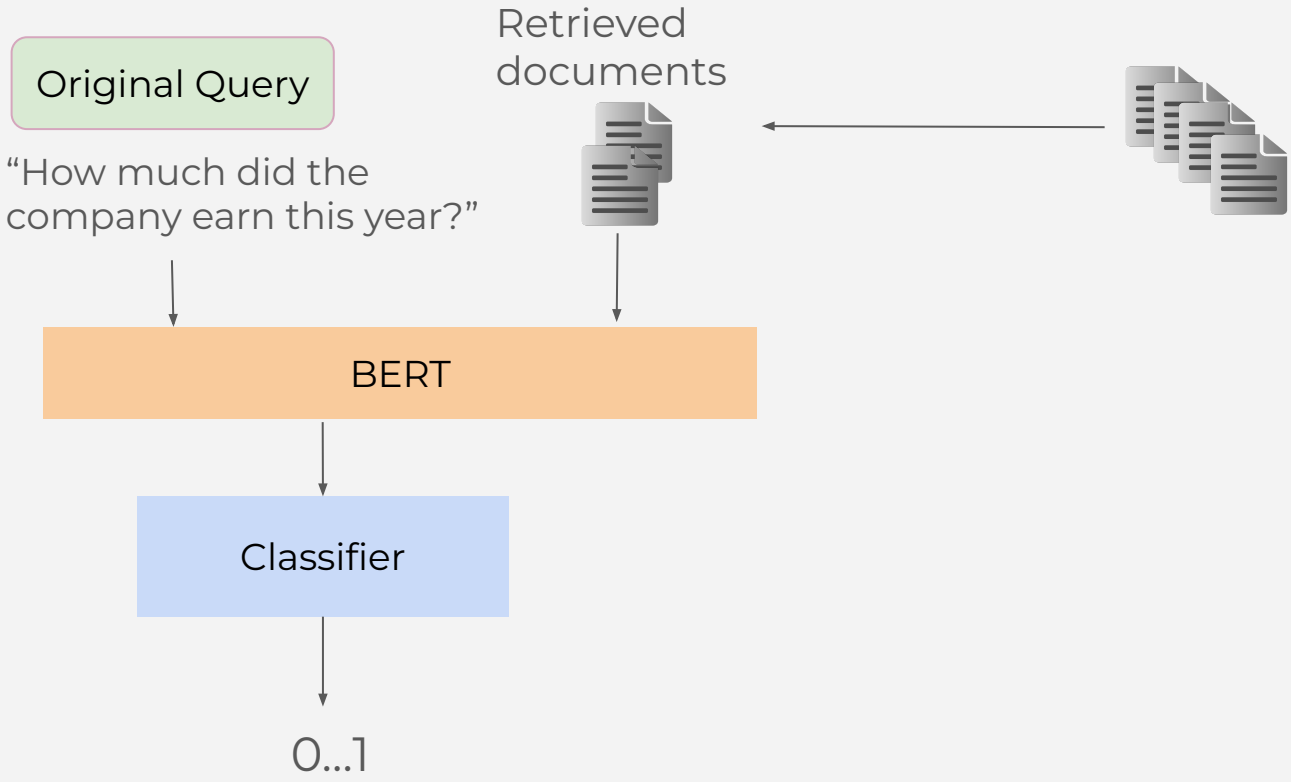
Re-ranking

Cross-Encoder Model

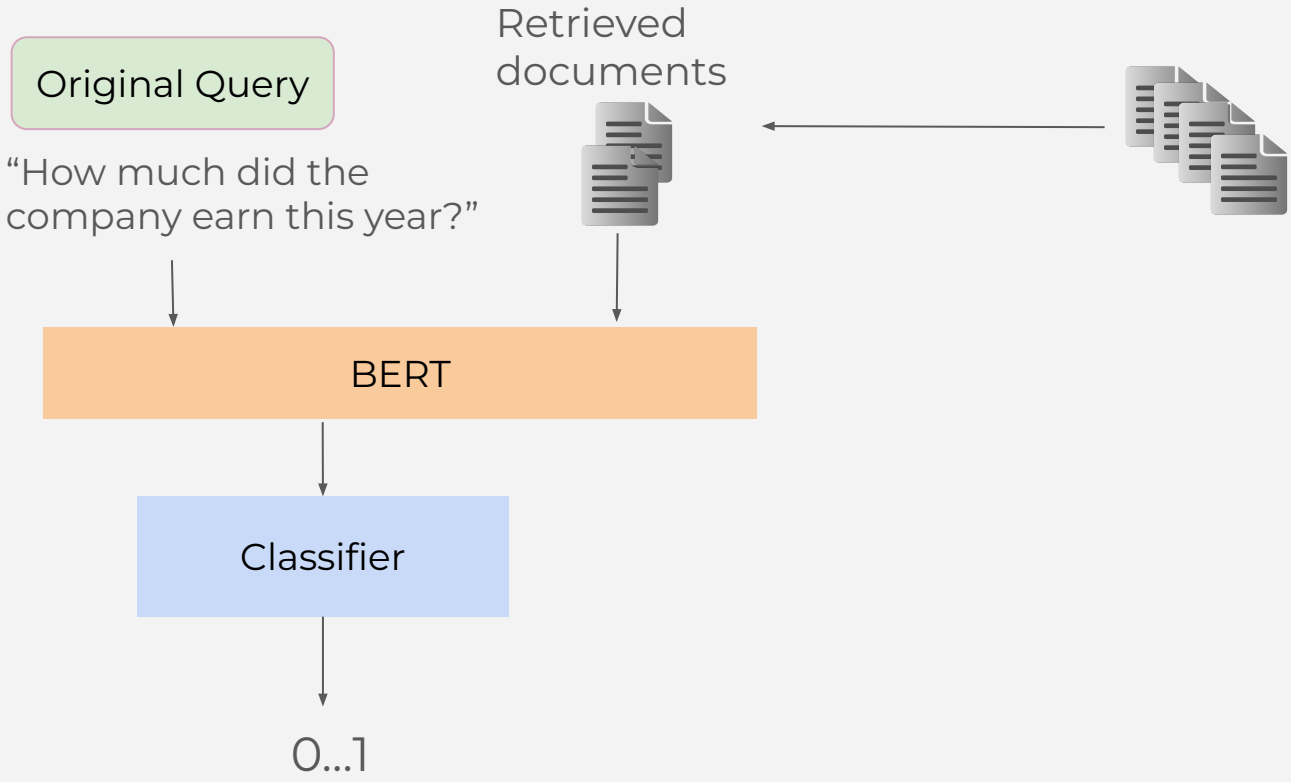




Cross-Encoder in Re-ranking

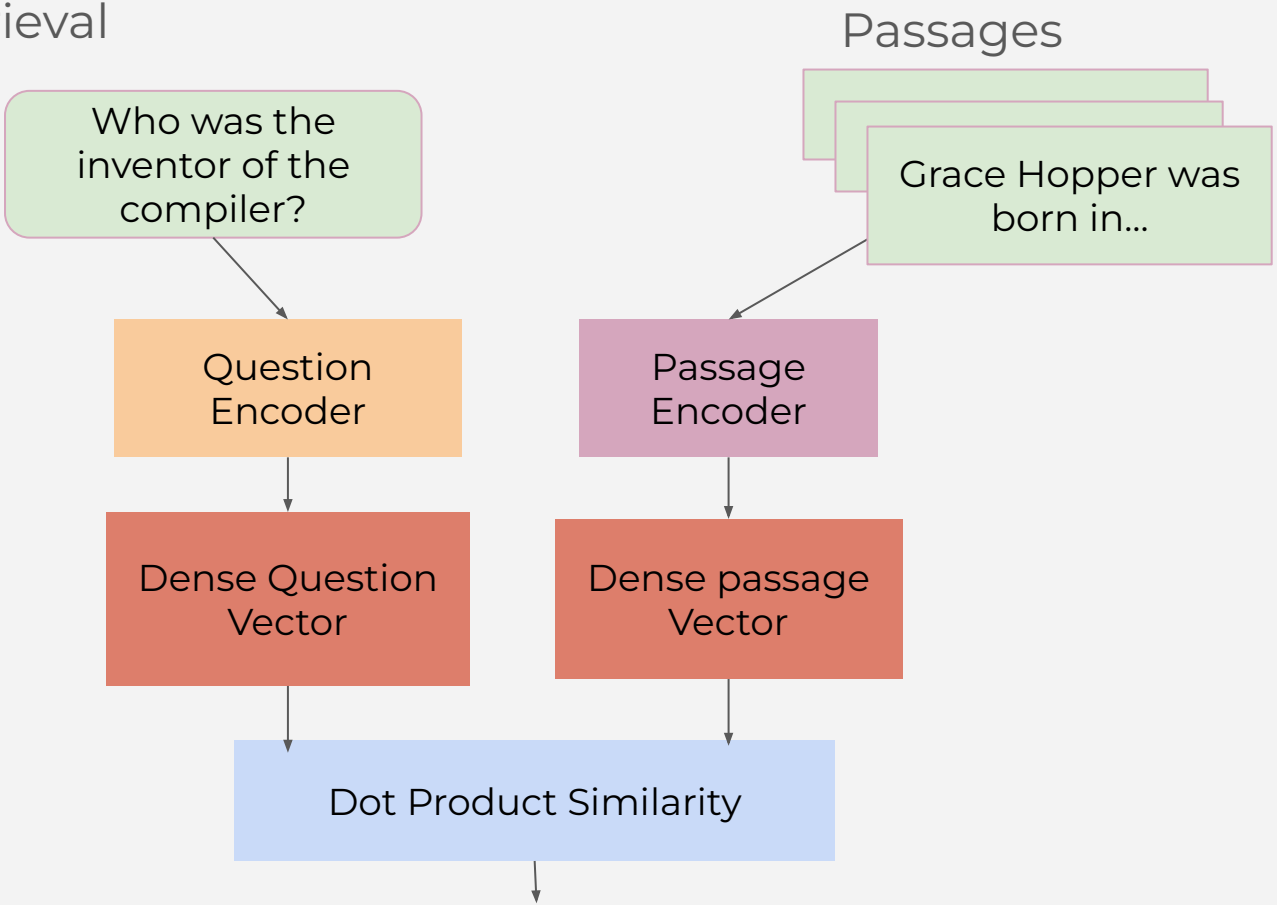


Cross-Encoder in Re-ranking



Other techniques...

Dense Passage Retrieval



Advanced RAG Techniques

DPR

Use cases:

- Open-Domain question answering
- Document retrieval
- Customer support

Advanced RAG Techniques

DPR

Hands-on

Other techniques...

~~Dense Passage Retrieval~~
Embedding Adaptors
Deep Chunking
RAG Fusion
...

Congratulations!

You made it to the end!

- Next steps...

Course Summary

- Advanced RAG Techniques
  - RAG overview
  - Naive RAG and Pitfalls
  - Deep dive into some RAG techniques
    - Query Expansion
      - *With generated answers*
      - *With multiple queries*
  - Re-ranking (Cross-encoder)
  - Dense Passage Retrieval (DPR)

Wrap up - Where to Go From Here?

- Keep learning
  - Get more ideas and build more applications and test different, less known Advanced RAG techniques

- Read more papers on RAG Techniques - https://arxiv.org/
- Challenge yourself to keep learning new skills!

Thank you!