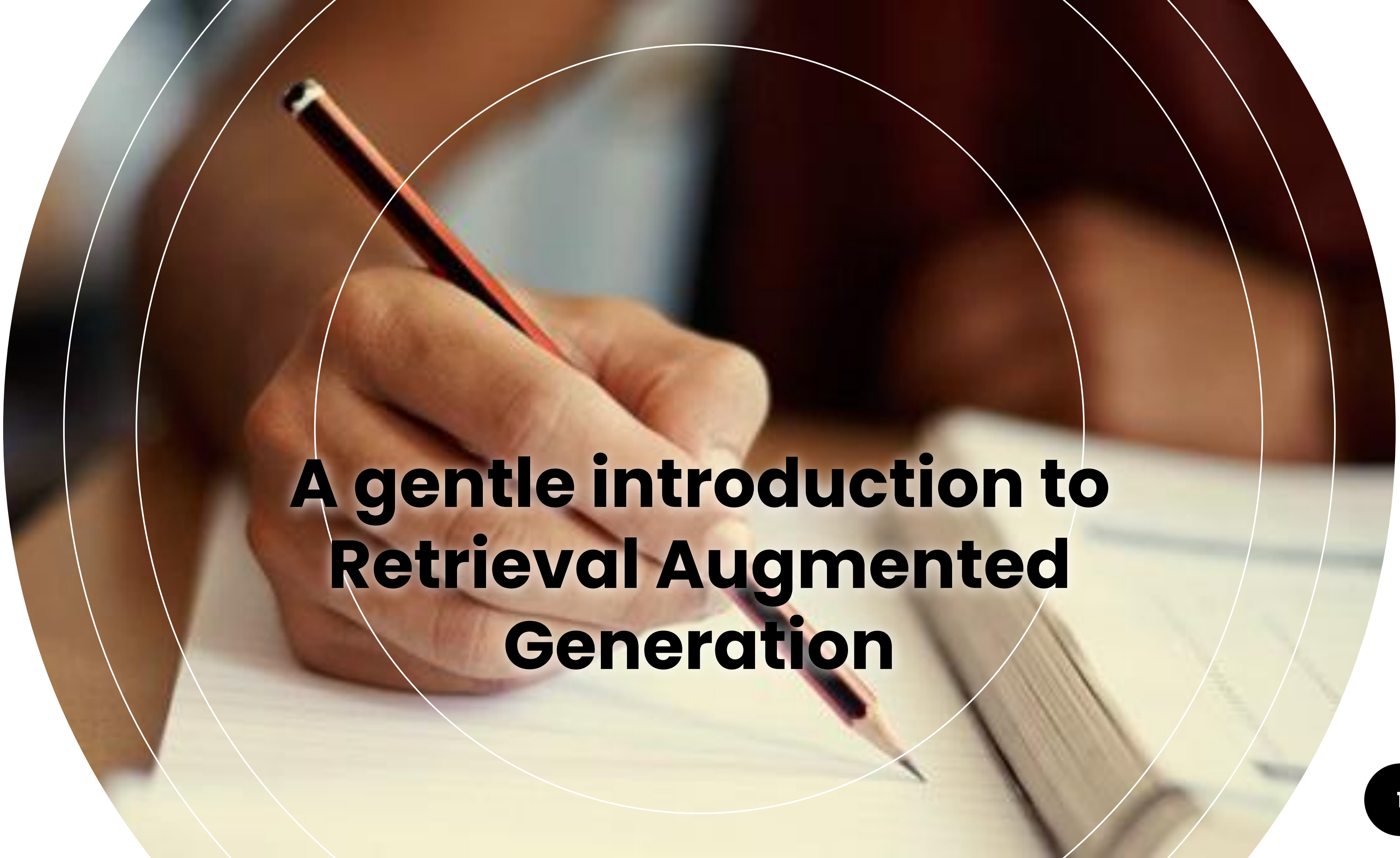
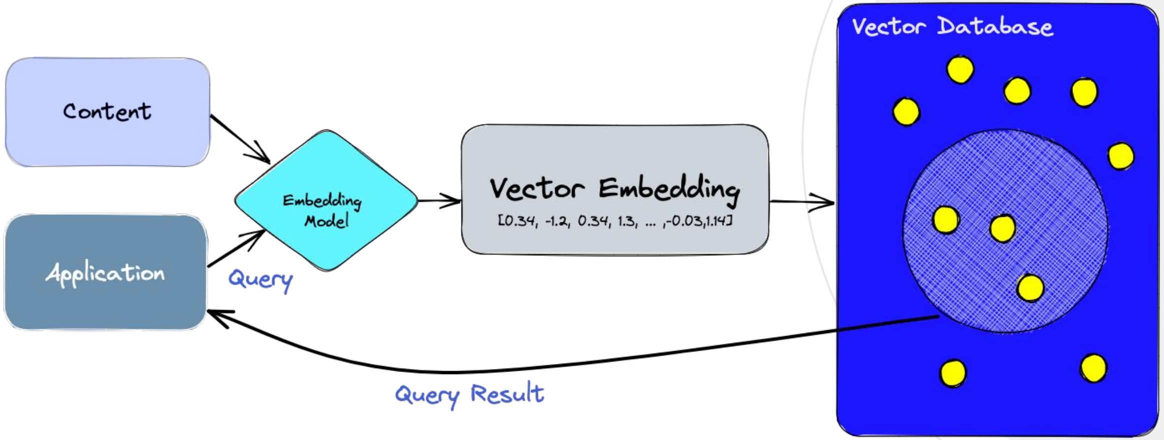


A gentle introduction to Retrieval Augmented Generation

Vector databases

- They make it possible to store texts in vector format
- For this purpose, an encoding mechanism is used
- Queries are encoded and the closest vectors are returned
- Can be associated with metadata

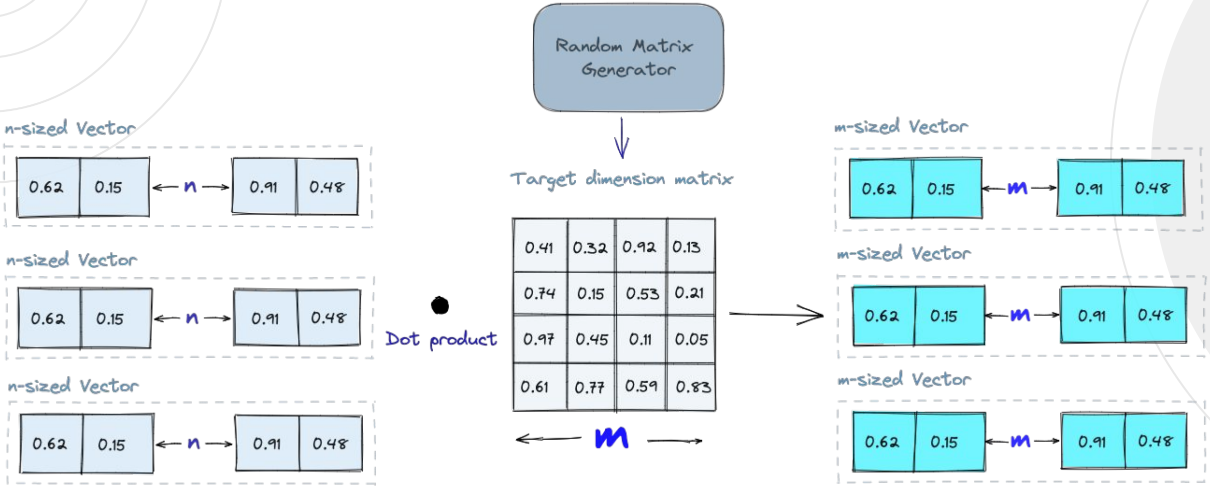
Vector databases



Source https://www.pinecone.io/learn/vector-database/

Vector databases

Indexing algorithms: *random projection*

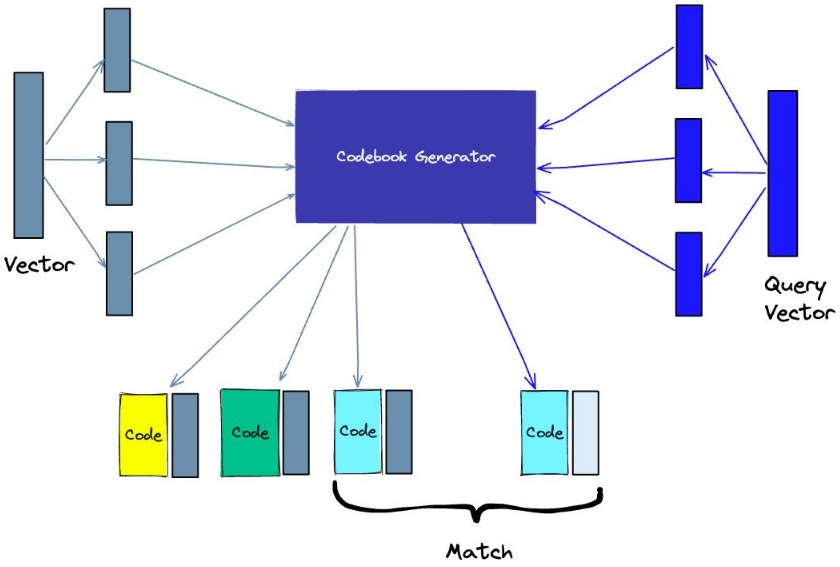


- The more random the projection matrix, the better
- Calculating this matrix can be costly

Source https://www.pinecone.io/learn/vector-database/

Vector databases

Indexing algorithms: *product quantization*

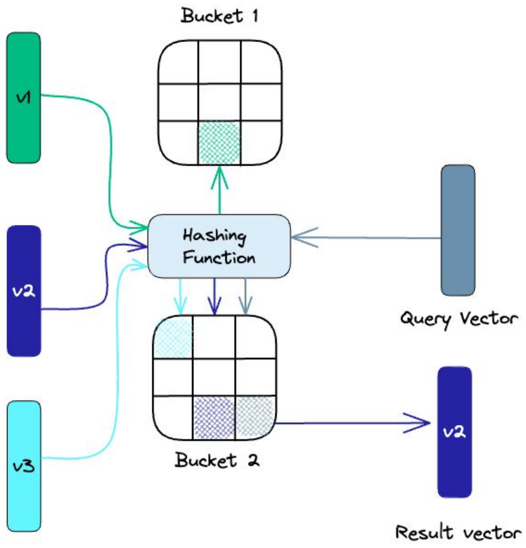


Source https://www.pinecone.io/learn/vector-database/

- We segment the vectors
- Clustering of each group of segments (the cluster of the first segments, the cluster of the second segments…)
- A code is assigned to each cluster

Vector databases

Indexing algorithms: *local sensitive hashing (LSH)*

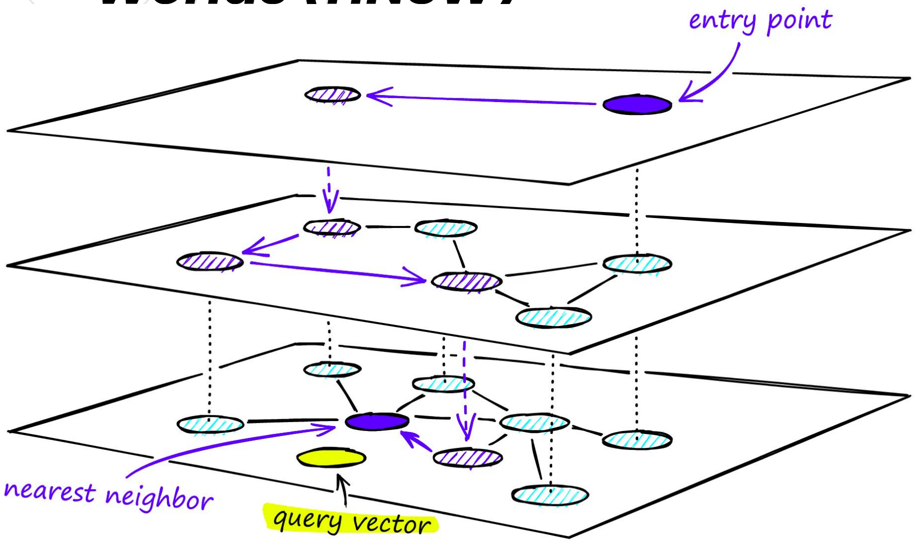


- A hash function separates groups of nearby vectors
- We only search in the block corresponding to the one returned by the hash function for the query

Source https://www.pinecone.io/learn/vector-database/

Vector databases

Indexing algorithms: *Hierarchical Navigable Small Worlds (HNSW)*

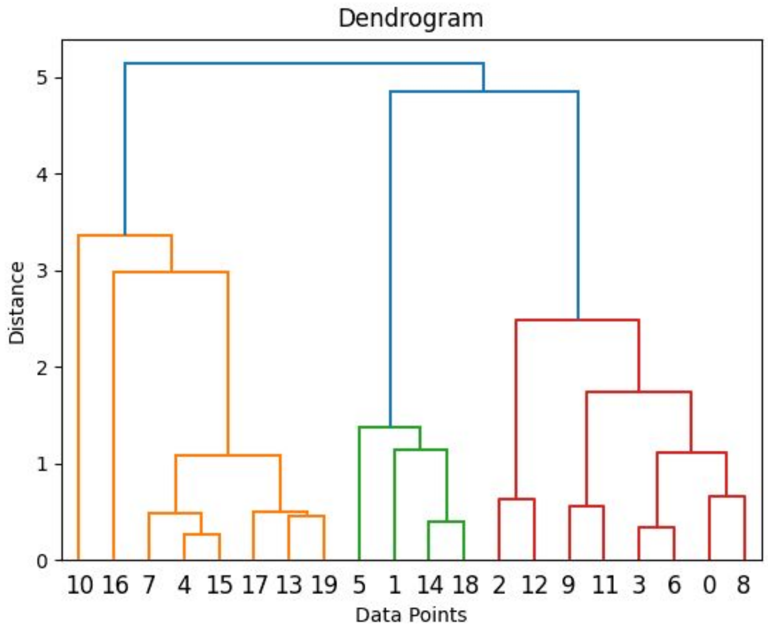


- Graph-based Approximate Nearest Neighbors algorithm
- State-of-the-art performance
- Vertices are linked based on proximity (closer vertices are linked)

Source https://www.pinecone.io/learn/series/faiss/hnsw/

Vector databases

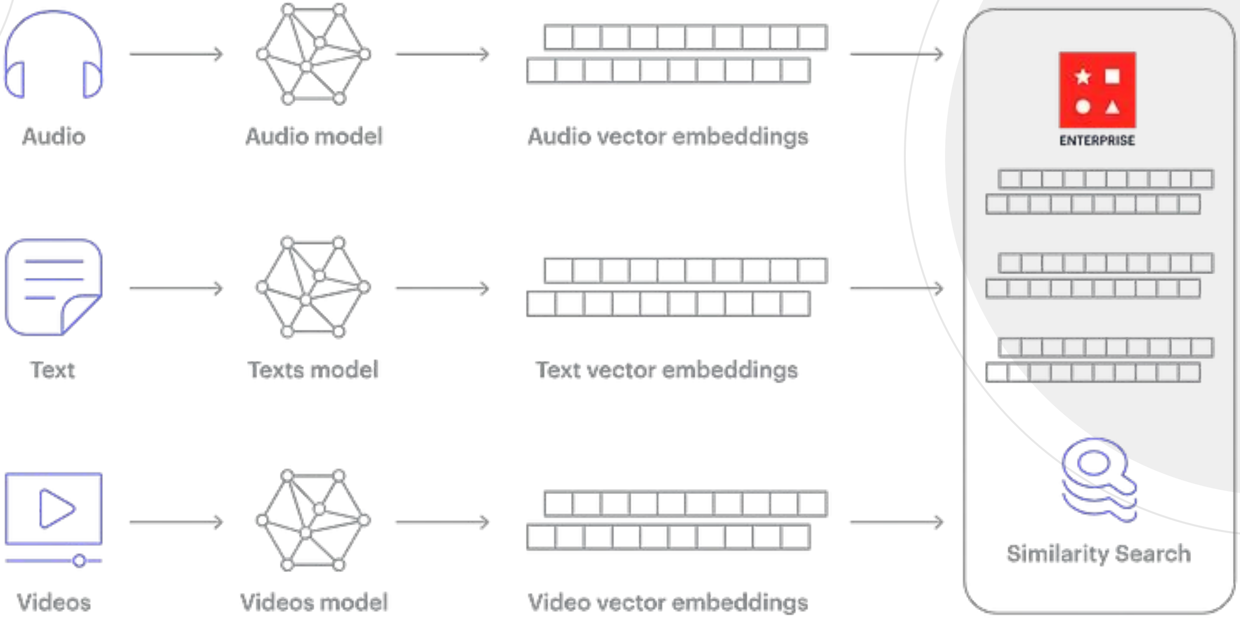
Indexing algorithms: *hierarchical clustering*



- We group texts by groups according to their distance from each other in a hierarchical manner.
- Each group is identified by its centroid
- We search by distance from the query vector to the centroids.

Vector databases

Everything vectorizable is indexable: multimodality



Vector databases

Examples

- Pinecone (proprietary)
- ChromaDB (open)
- QDrant
- Redis
- ...

Pinecone

Chroma

Language Models

Language Model: Definition

Definition

A **language model** computes a probability distribution over words from sequences of words.

$$P(w_m \mid w_1, w_2, ..., w_{m-1})$$

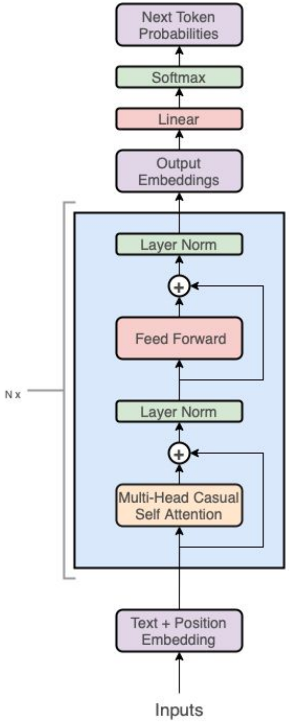
```
P(“chips” | “fish and”) >> P(“milk” | “fish and”)
```

Language Model: Types

Types

- Hidden Markov Models
- N-gram
- Statistical, probabilistic
- Recurrent neural networks
- Transformer decoders

Language Model: Transformer



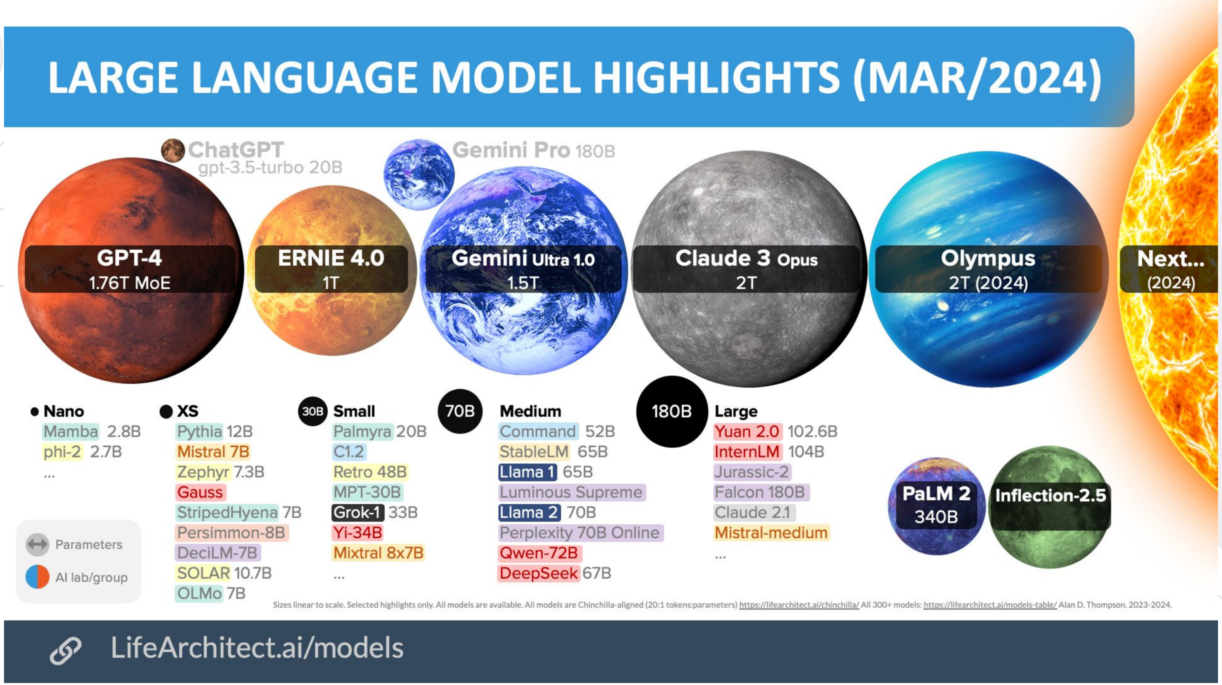
$$\text{softmax}(x)_i = \frac{e^{x_i}}{\sum_j e^{x_j}}$$

$$\text{LayerNorm}(x) = \gamma \cdot \frac{x - \mu}{\sqrt{\sigma^2}} + \beta$$

$$\begin{aligned}\text{GELU}(x) &= xP(X \le x) = x\Phi(x) \\ &\approx 0.5x\left(1 + \tanh\left[\sqrt{2/\pi}\left(x + 0.044715x^3\right)\right]\right)\end{aligned}$$

$$\text{attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Language Model: LARGE



- \+ Llama3
- \+ Phi3

...

Source https://lifearchitect.substack.com/p/the-memo-8mar2024

More info [Stanford Report on AI 2024]()

Language Model: Problems

- Computationally expensive
- Hallucinations
- Limitations in reasoning
    - Mathematical operations
    - Logic
    - Planning
- Limit of input sequence (context)
- Learning from experience
- Bias
- Privacy
- Security

Retrieval Augmented Generation

- A **true information retrieval system** (and not a document retriever)
- Queries in **natural language** and gives answers in natural language.
- Use of a document base as a **knowledge base**
- We can "talk" with texts
- Knowledge is NOT in the model

Retrieval Augmented Generation

ChatGPT

Regarding your new question about the Fest in Geneva, I will search for current information on this. One moment, please.

Se ha buscado 6 sitios

The Geneva Fest, also known as "Fêtes de Genève," will take place from August 22 to August 24, 2024. This vibrant event will feature a range of activities and performances, celebrating the rich cultural tapestry of Geneva (Songkick).

Additionally, the Montreux Jazz Festival, a nearby prominent event, will be held from July 5 to July 20, 2024, offering a lineup of various renowned artists (Music Festival Wizard).

For more information on specific events and schedules, you can visit the official websites or event pages.

Get citation

When is this year Geneva Fest?

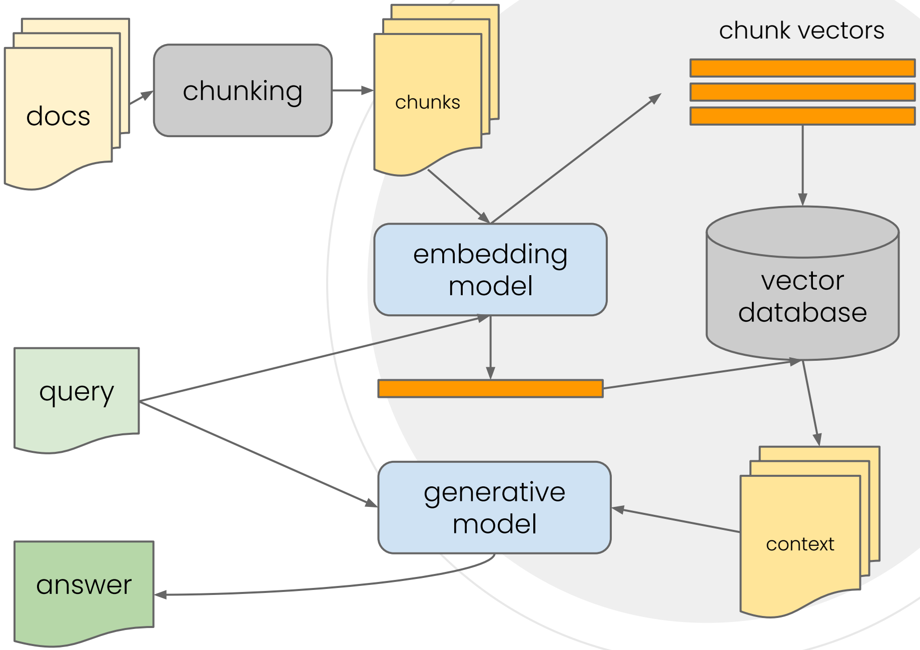
RAG: architecture

Components:

- Document collection
- Encoder model
- Vector database
- Language model

Phases:

- Loading
- Indexing
- Storage
- Query (direct, step-by-step...)
- Evaluation



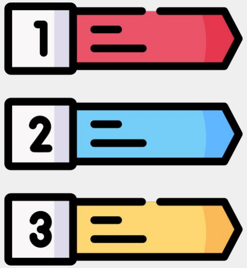
RAG: loading

Data sources:

- Existing collections
- Databases
- Information on the web
- PDF or other formats

RAG: indexing

- Chunk size
- Indexing algorithm: LSH, hierarchical...
- Encoding model
    - BGE-Large
    - E5 ...



RAG: query

- From the chunks (context), the generative LLM must produce a response.
- This involves:
    - Retrieval
    - Generation

RAG: query

Retrieval

We can have different collections and different strategies for each one of them

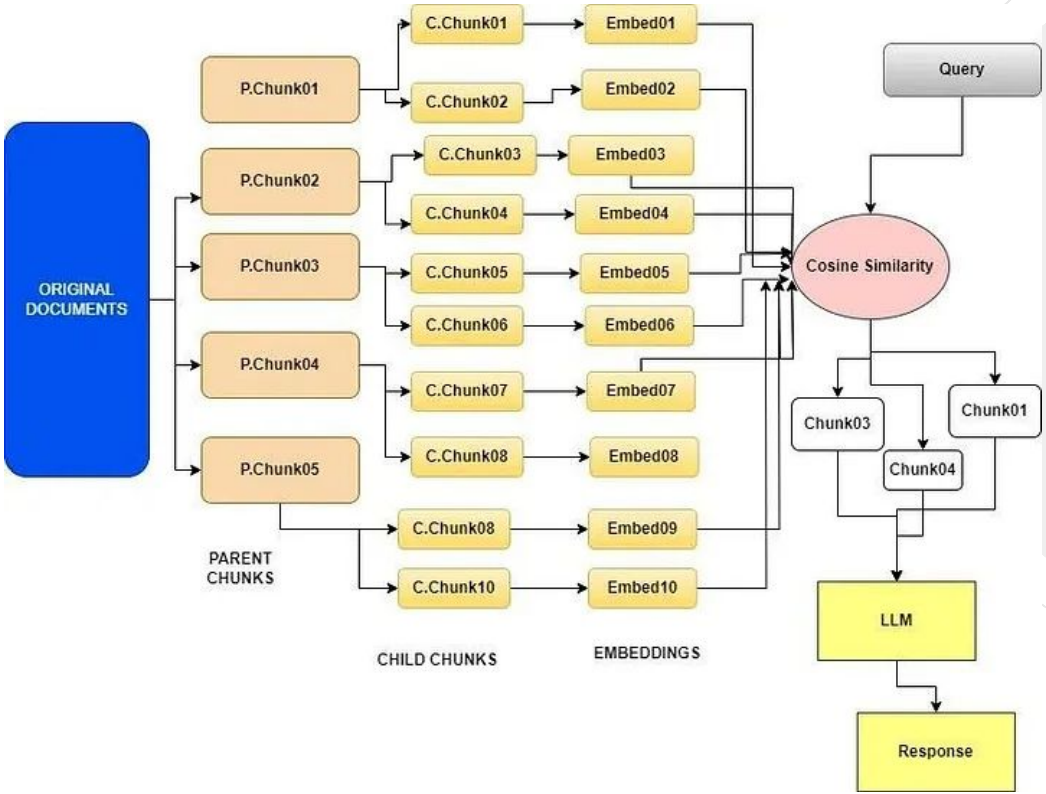
- Hybrid with BM25
- Simple merging
- Reciprocal reranking
- Composite
- Auto-retrieval (LLM to select metadata)
- Translation (text → SQL)
- Knowledge networks...

$$RRFscore(d \in D) = \sum_{r \in R} \frac{1}{k + r(d)}; k = 60$$

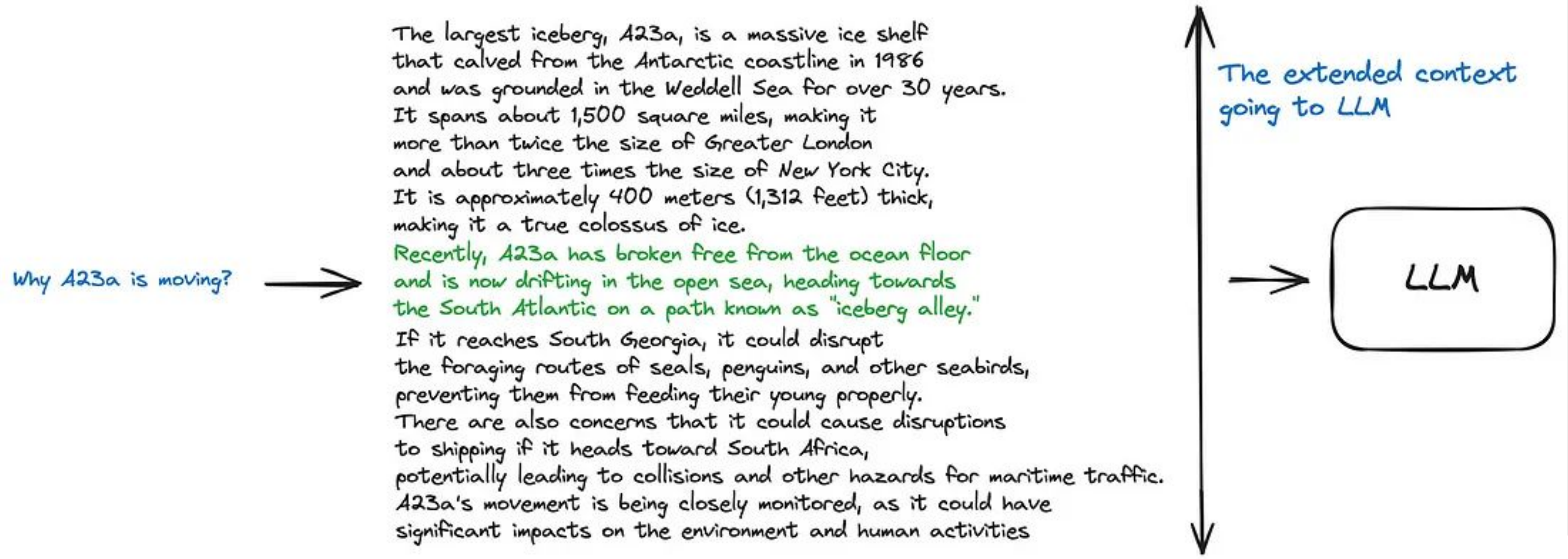
RAG: retrieval strategies

- Hierarchical chunks
- Context window
- Hybridization with classic RI

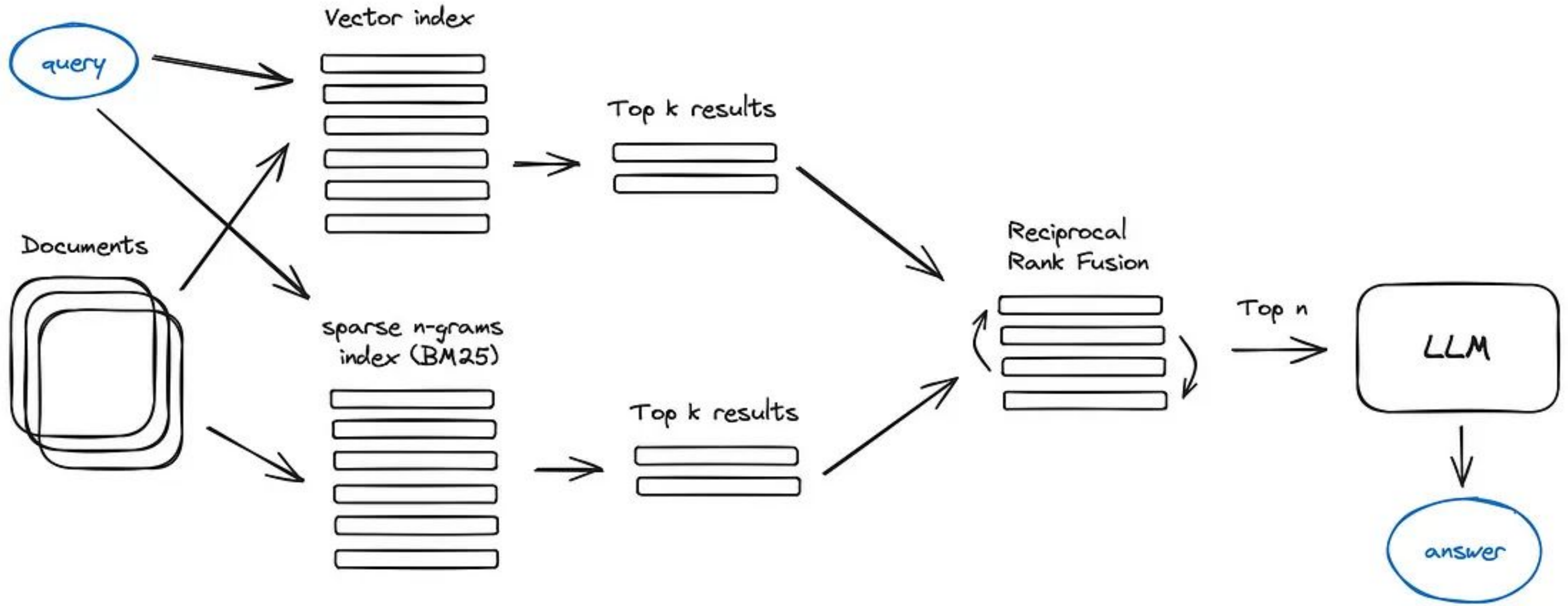
RAG: hierarchical chunks



RAG: context window



RAG: hybridization with classic RI



RAG: query

Recovery: chunk processing

Once the chunks (also called nodes) have been retrieved, we can process them:

- Filter by threshold
- Filtering by keywords (terms that SHOULD appear or terms that SHOULD NOT appear)
- Node augmentation: augment the contexts based on the relationships between the nodes

RAG: query

Response synthesis

- Prompt engineering
  - Node by node and then summarize
  - Concatenation
  - Reasoning chain
  - Reasoning tree
  - …
- It is important to know well
  - The LLM used to ensure that it responds in the right language and format.
  - The nature of the documents and types of queries

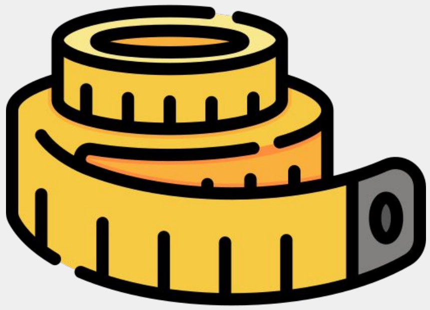
RAG: evaluation

Retrieval evaluation

- Context precision
- Context recall

Generation evaluation

- Faithfulness
- Answer relevance



RAG: evaluation

Context precision

We need: query + contexts + reference answer

$$\text{Context Precision@K} = \frac{\sum_{k=1}^{K} (\text{Precision@k} \times v_k)}{\text{Total number of relevant items in the top } K \text{ results}}$$

$$\text{Precision@k} = \frac{\text{true positives@k}}{(\text{true positives@k} + \text{false positives@k})}$$

Where K is the total number of chunks in `contexts` and $v_k \in \{0, 1\}$ is the relevance indicator at rank k.

RAG: evaluation

Context precision: example

Hint

Question: Where is France and what is it's capital? Ground truth: France is in Western Europe and its capital is Paris.

High context precision: ["France, in Western Europe, encompasses medieval cities, alpine villages and Mediterranean beaches. Paris, its capital, is famed for its fashion houses, classical art museums including the Louvre and monuments like the Eiffel Tower", "The country is also renowned for its wines and sophisticated cuisine. Lascaux's ancient cave drawings, Lyon's Roman theater and the vast Palace of Versailles attest to its rich history."]

Low context precision: ["The country is also renowned for its wines and sophisticated cuisine. Lascaux's ancient cave drawings, Lyon's Roman theater and", "France, in Western Europe, encompasses medieval cities, alpine villages and Mediterranean beaches. Paris, its capital, is famed for its fashion houses, classical art museums including the Louvre and monuments like the Eiffel Tower",]

CP@2 = (1+0.5)/2 = **0.75**

$$\text{Precision@1} = \frac{0}{1} = 0$$

$$\text{Precision@2} = \frac{1}{2} = 0.5$$

CP@2 = (0+0.5)/2 = **0.25**

RAG: evaluation

Context recall

We need: contexts + ground truth answers

$$\text{context recall} = \frac{|\text{GT sentences that can be attributed to context}|}{|\text{Number of sentences in GT}|}$$

RAG: evaluation

Context recall: example

Hint

Question: Where is France and what is it's capital?

Ground truth: France is in Western Europe and its capital is Paris.

High context recall: France, in Western Europe, encompasses medieval cities, alpine villages and Mediterranean beaches. Paris, its capital, is famed for its fashion houses, classical art museums including the Louvre and monuments like the Eiffel Tower.

Low context recall: France, in Western Europe, encompasses medieval cities, alpine villages and Mediterranean beaches. The country is also renowned for its wines and sophisticated cuisine. Lascaux's ancient cave drawings, Lyon's Roman theater and the vast Palace of Versailles attest to its rich history.

$CR=\frac{1}{2}=0.5$

RAG: evaluation

Other retrieval metrics:

- Context relevancy

$$\text{context relevancy} = \frac{|S|}{|\text{Total number of sentences in retrieved context}|}$$

S: relevant chunks

- Context entity recall

$$\text{context entity recall} = \frac{|CE \cap GE|}{|GE|}$$

CE: Context entities

GE: Ground truth entities

RAG: evaluation

Faithfulness (generation evaluation)

- An LLM is used to extract claims
- We are NOT looking for the *veracity* of the claims, but whether the LLM has responded based on the context.

$$\text{Faithfulness score} = \frac{|\text{Number of claims in the generated answer that can be inferred from given context}|}{|\text{Total number of claims in the generated answer}|}$$

RAG: evaluation

Faithfulness: example

Hint

Question: Where and when was Einstein born?

Context: Albert Einstein (born 14 March 1879) was a German-born theoretical physicist, widely held to be one of the greatest and most influential scientists of all time

High faithfulness answer: Einstein was born in Germany on 14th March 1879.

Low faithfulness answer: Einstein was born in Germany on 20th March 1879.

RAG: evaluation

Answer relevancy

- It is calculated by using *embeddings*
- New queries generated from the original query
- We need: query + context + answer

$$\text{answer relevancy} = \frac{1}{N}\sum_{i=1}^{N} cos(E_{g_i}, E_o)$$

$$\text{answer relevancy} = \frac{1}{N}\sum_{i=1}^{N} \frac{E_{g_i} \cdot E_o}{\|E_{g_i}\|\|E_o\|}$$

RAG: evaluation

Answer relevancy: example

Hint

Question: Where is France and what is it's capital?

Low relevance answer: France is in western Europe.

High relevance answer: France is in western Europe and Paris is its capital.

RAG: examples

RAG development systems

- LlamaIndex
- LangChain
- Haystack
- OpenAI Assistants

Evaluation

- RAGAS

Summary

- Deep neural networks are revolutionizing information retrieval systems
- Deep neural networks are experiencing unparalleled growth.
- There are many challenges: optimization, cost, control of the generated response, biases, out-of-context questions, personalization, etc.
- RAGs are evolving into multi-agent systems.