Large Language Models with Semantic Search

Keyword Search

Query

What color is the grass

Responses

Tomorrow is Saturday

The grass is green

The capital of Canada is Ottawa

The sky is blue

A whale is a mammal

Keyword Search

Query

What color is the grass

Responses	**Number of common words**
Tomorrow **is** Saturday	1
The grass is green	3
The capital of Canada **is** Ottawa	2
The sky **is** blue	2
A whale **is** a mammal	1

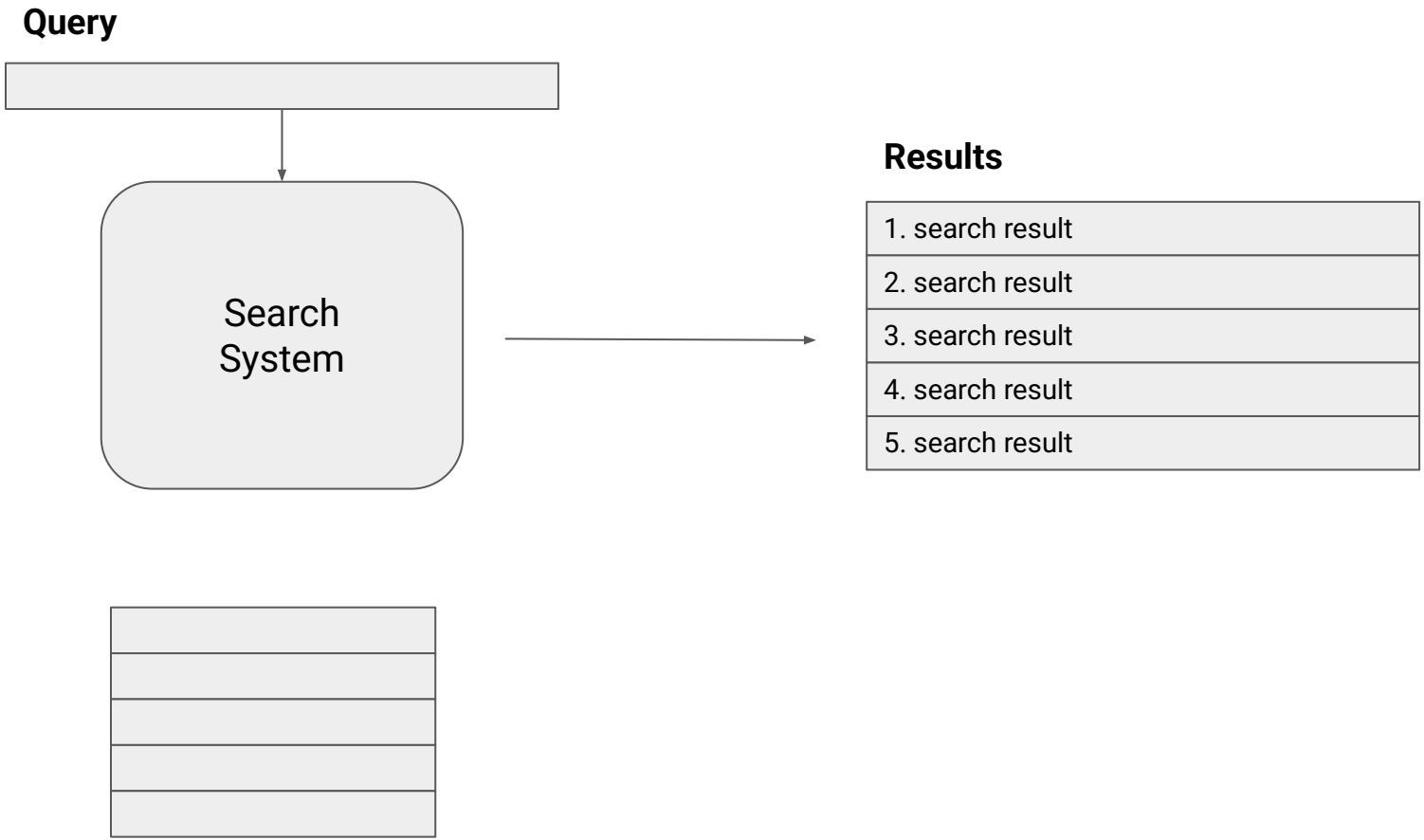
Keyword Search

Query

What color is the grass

Responses	**Number of common words**
Tomorrow **is** Saturday	1
The grass is green	3
The capital of Canada **is** Ottawa	2
The sky **is** blue	2
A whale **is** a mammal	1

Search at a High Level



Keyword Search Internals

Query

What color is the sky?

Search System

Retrieval
First Stage

Reranking
Second Stage

BM25 Algorithm

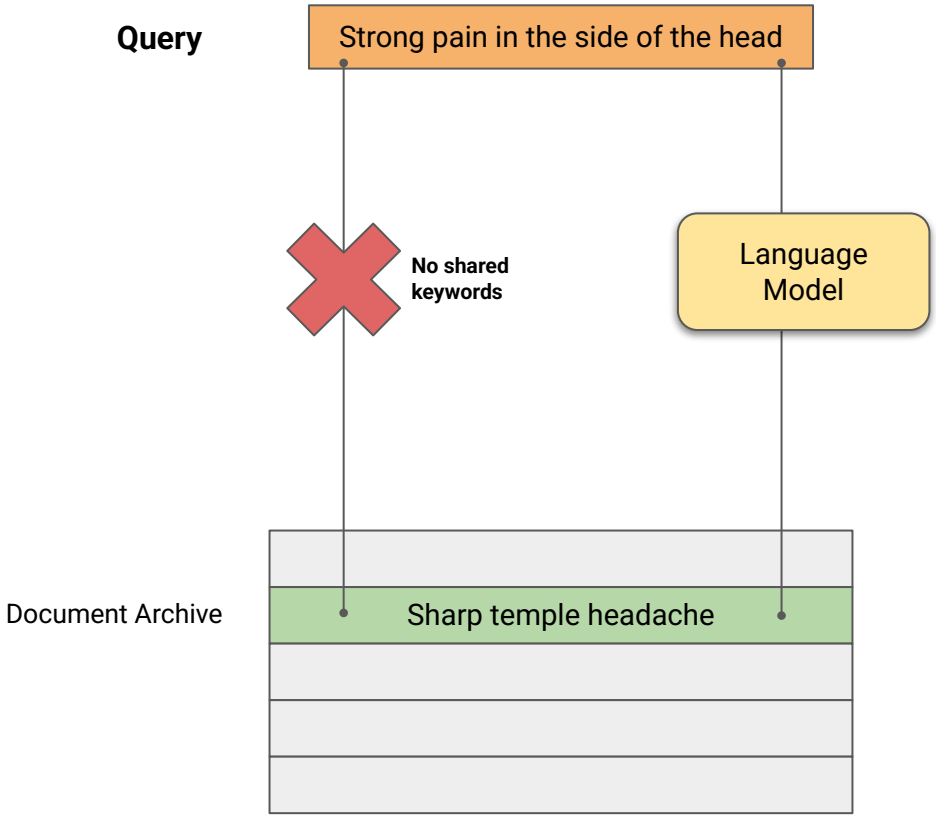
Results

1. search result
2. search result
3. search result
4. search result
5. search result

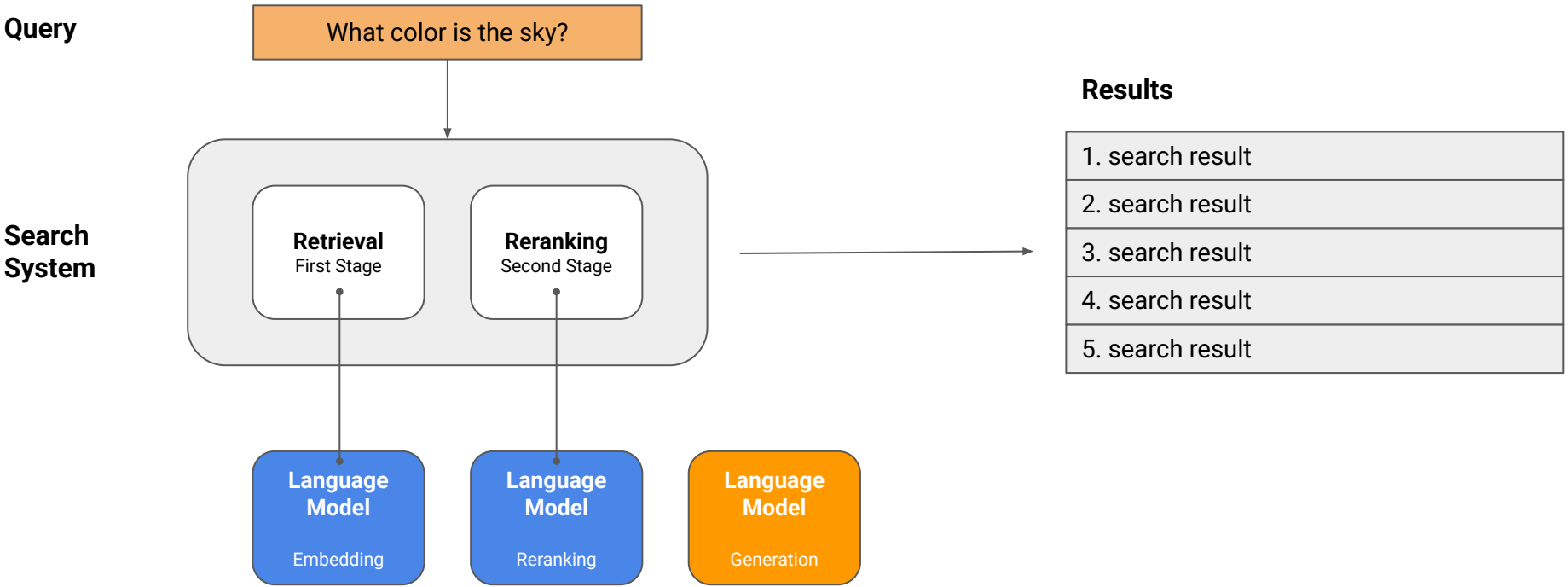
Inverted Index

keyword	Document IDs
abacus	1, 38, 18
- - -	
Color	23, 804
- - -	
Sky	804, 922

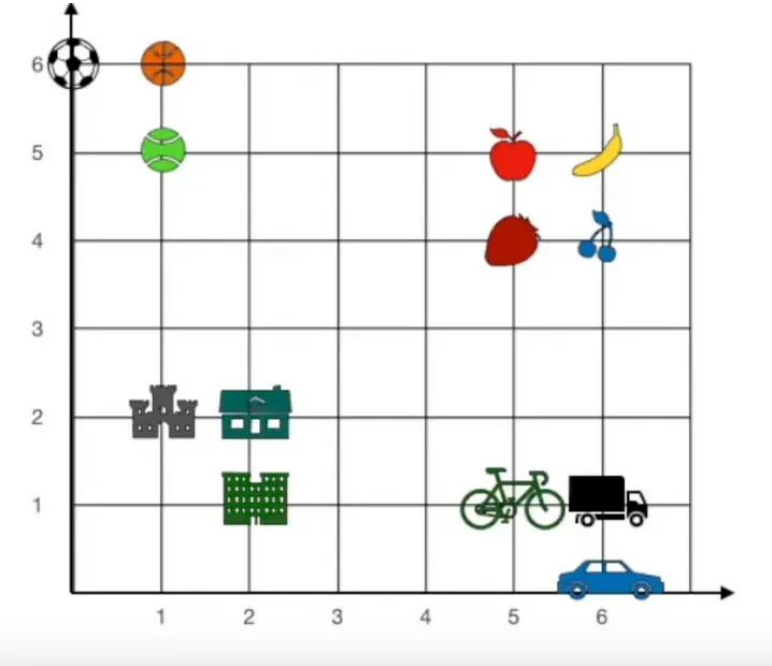
Limitation of Keyword Matching



Language Models can Improve both Search Stages



Embeddings



Word	Numbers	
Apple	5	5
Banana	6	5
Strawberry	5	4
Cherry	6	4
Soccer	0	6
Basketball	1	6
Tennis	1	5
Castle	1	2
House	2	2
Building	2	1
Bicycle	5	1
Truck	6	1
Car	6	0

Word Embeddings

Word	Numbers	
Apple	5	5
Soccer	0	6
House	2	2
Car	6	0

Word	Numbers			
A	-0.82	-0.32	- - -	0.23
I am going to school today	0.42	1.28	- - -	-0.06
- - - -	- - -	- - -	- - -	- - -
Zygote	-0.74	-1.02	- - -	1.35

Text Embeddings

Text	Numbers				
Hello, how are you?	0.39	0.49	- - -	-1.01	-0.72
I am going to school today	-0.79	-0.05	- - -	-0.94	2.71
- - - -	- - -	- - -	- - -	- - -	- - -
Once upon a time	3.23	-0.23	- - -	-1.45	0.82
Hi, how is it going?	0.41	0.48	- - -	-0.98	-0.66

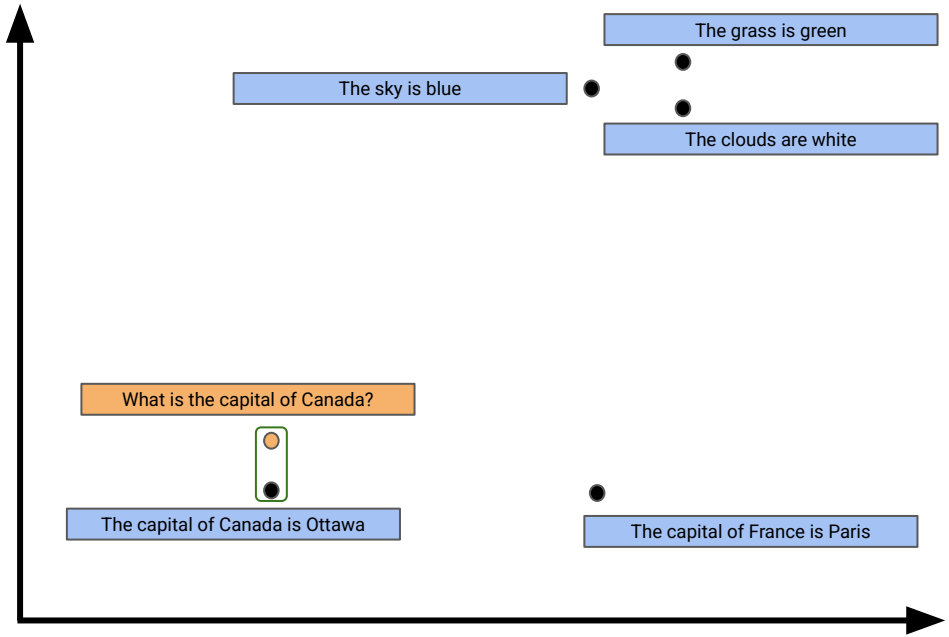
Dense Retrieval

Query

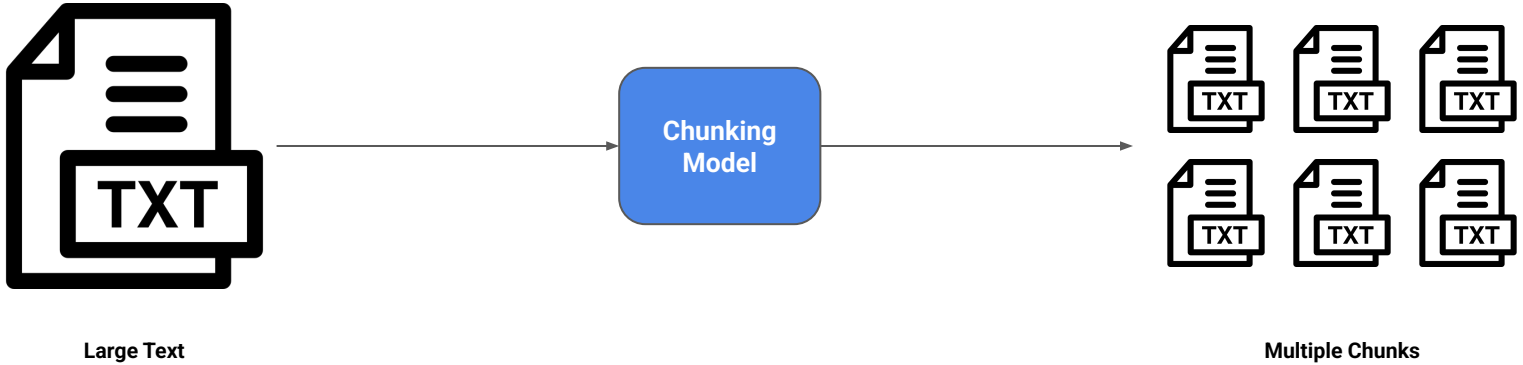
What is the capital of Canada

Responses

The capital of Canada is Ottawa
Toronto is in Canada
The sky is blue
The grass is green
The clouds are white



Chunking



Large Text

Multiple Chunks

ANN Vector Search vs Vector Databases

Approximate Nearest-Neighbor Vector search libraries

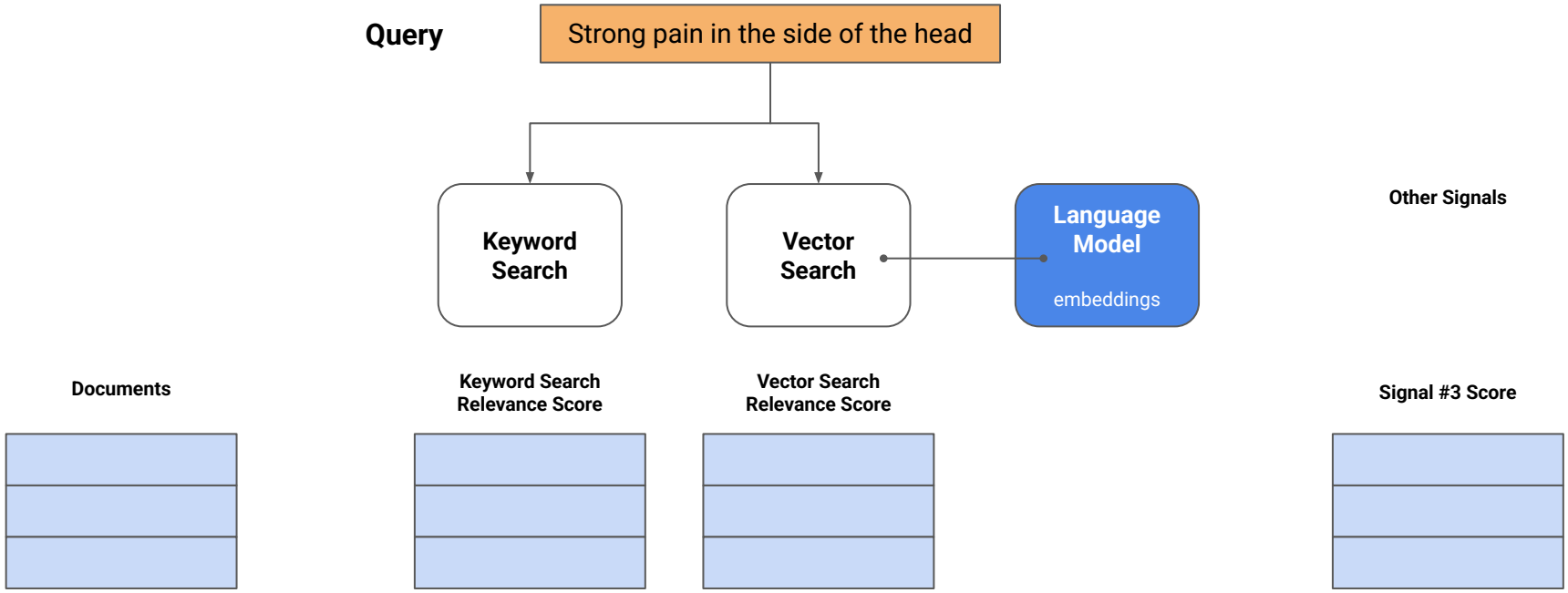
- Annoy
- FAISS
- ScaNN

- Easy to set up
- Store vectors only

Vector Databases

- Store vectors and text
- Easier to update (add new records)
- Allow filtering and more advanced queries

Hybrid Search: Keyword + Vector



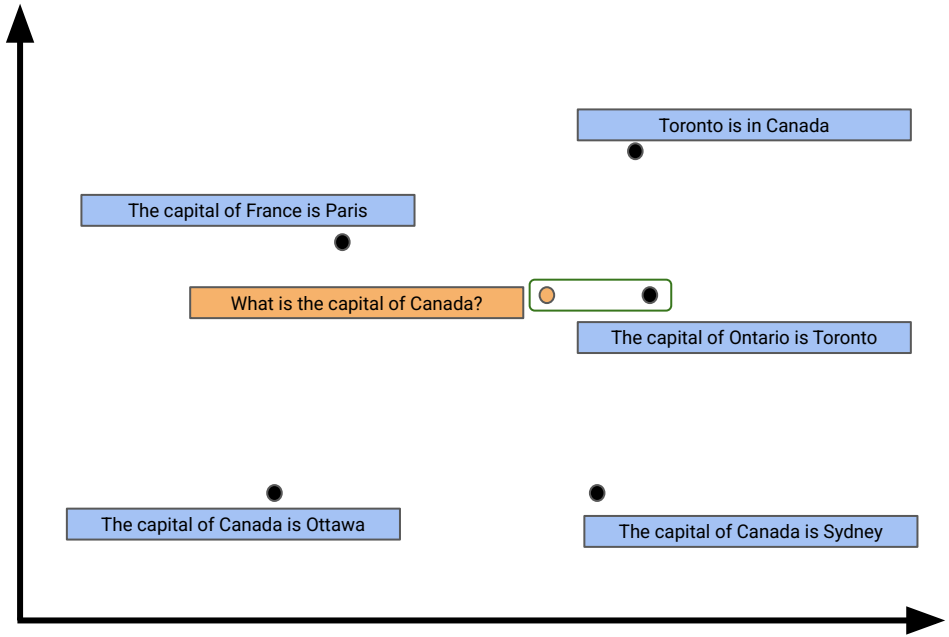
Dense Retrieval is also not Perfect

Query

What is the capital of Canada

Responses

- The capital of Canada is Ottawa
- Toronto is in Canada
- The capital of France is Paris
- The capital of Canada is Sydney
- The capital of Ontario is Toronto



Solution: ReRank

Query

What is the capital of Canada

Top Responses

Europe is a continent
The capital of France is Paris
The grass is green
The sky is blue
Toronto is in Canada
Tomorrow is Sunday
The capital of Canada is Ottawa
The capital of Canada is Sydney
Most apples are red
The capital of Ontario is Toronto

The capital of France is Paris
Toronto is in Canada
The capital of Canada is Ottawa
The capital of Canada is Sydney
The capital of Ontario is Toronto

Relevance

0.2
0.3
0.9
0.6
0.5

ReRank is Trained on

Many queries with correct answers:

What is the capital of Canada	The capital of Canada is Ottawa
What is the capital of France	The capital of France is Paris
⋮	⋮
What color is the sky?	The sky is blue

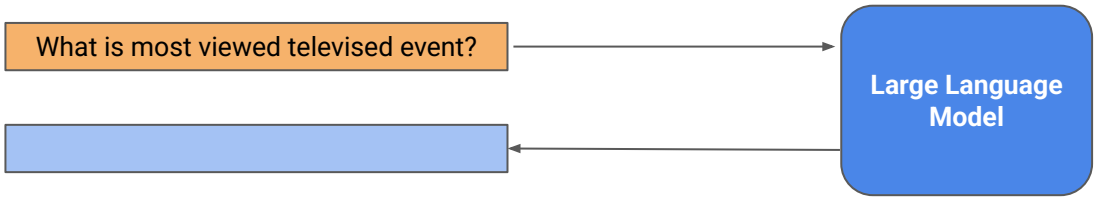
Many queries with wrong answers:

What is the capital of Canada	Toronto is in Canada
What is the capital of France	The capital of France is Barcelona
⋮	⋮
What color is the sky?	The sky is red

Evaluating Search Systems

- **Mean Average Precision (MAP)**
- **Mean Reciprocal Rank (MRR)**
- **Normalized Discounted Cumulative Gain (NDCG)**

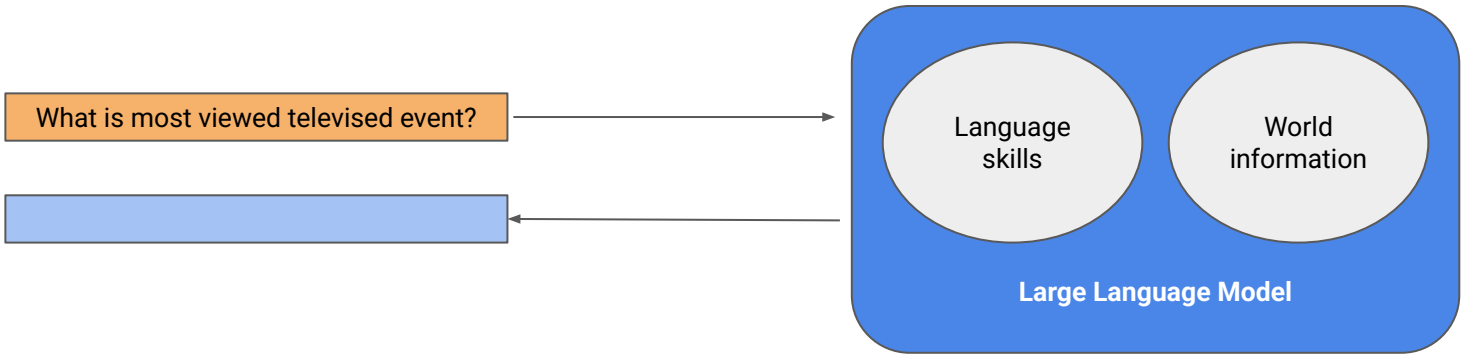
Search can help LLMs in Multiple ways



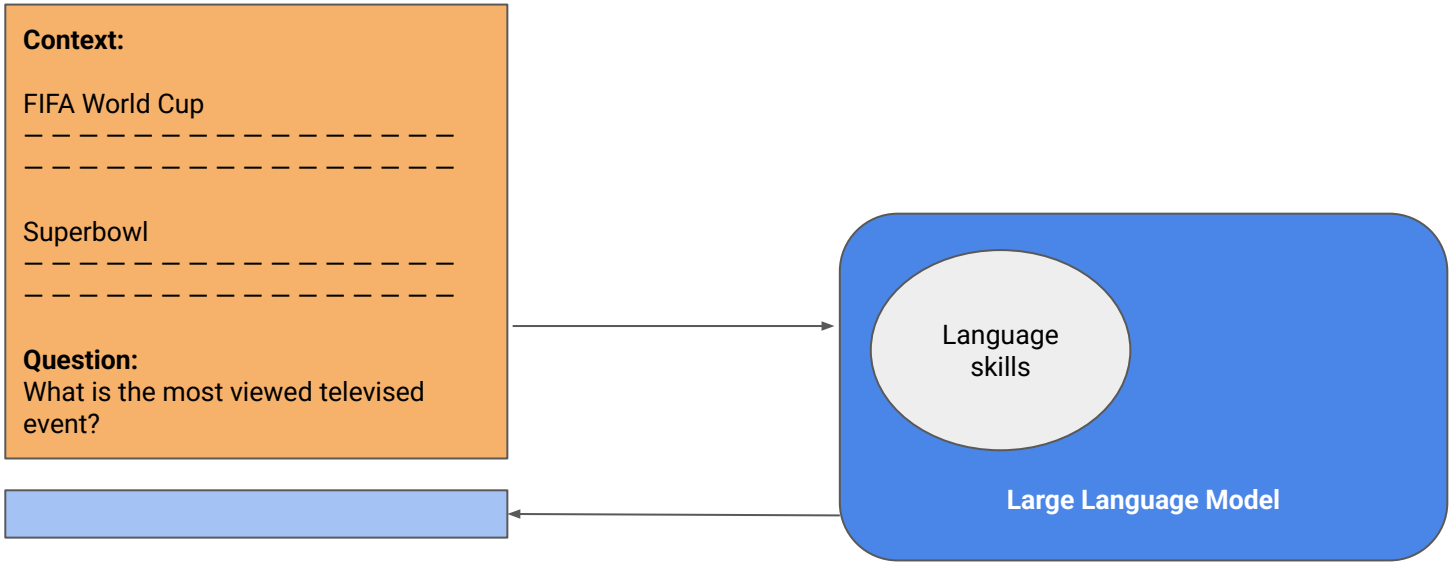
Retrieval can help LLMs with:

- Mean Average Precision (MAP)
- Mean Reciprocal Rank (MRR)
- Normalized Discounted Cumulative Gain (NDCG)

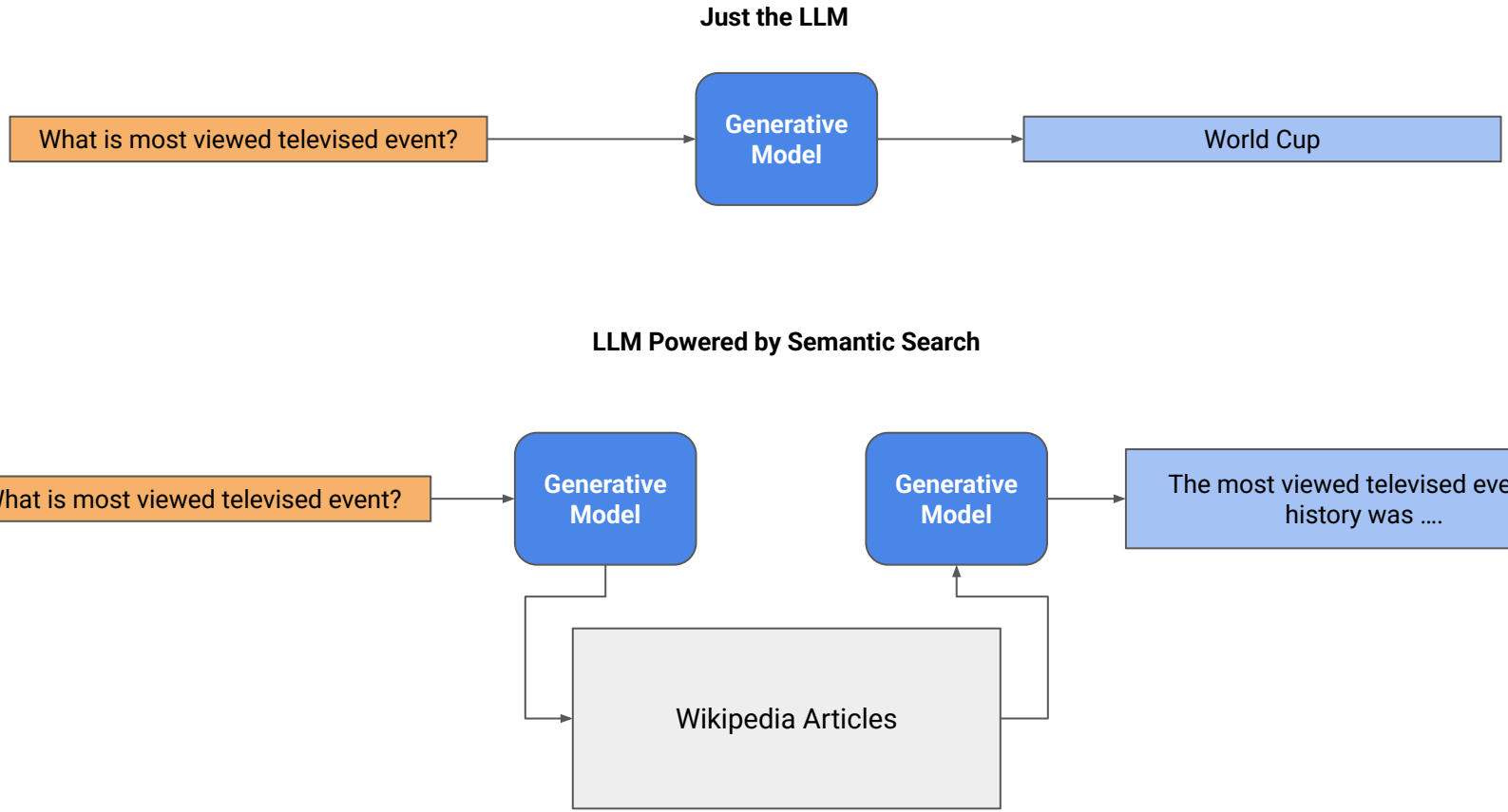
Where is the Information Stored?



Search can add some Context



Generating Answers



THANK YOU