Unlock NextGen Product Search with ML and LLM Innovations

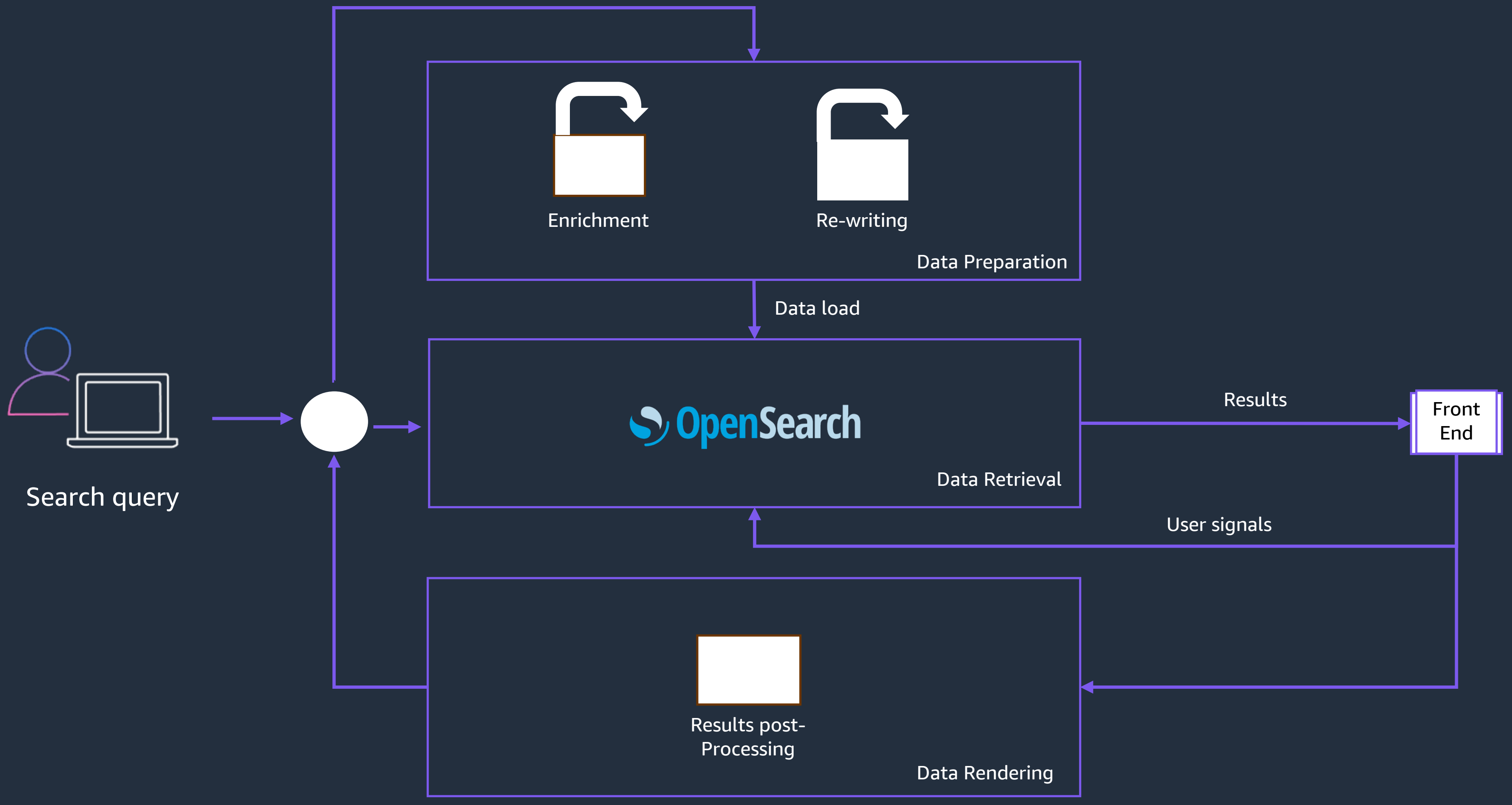
Praveen Mohan Prasad

Analytics Specialist TAM
AWS

Hajer Bouafif

Sr. WW Solutions Architect - OpenSearch
AWS

Search lifecycle



OpenSearch Growth

Apache 2.0 License
Linux Foundation

>700MM
OpenSearch project downloads since launch in Q3 2021

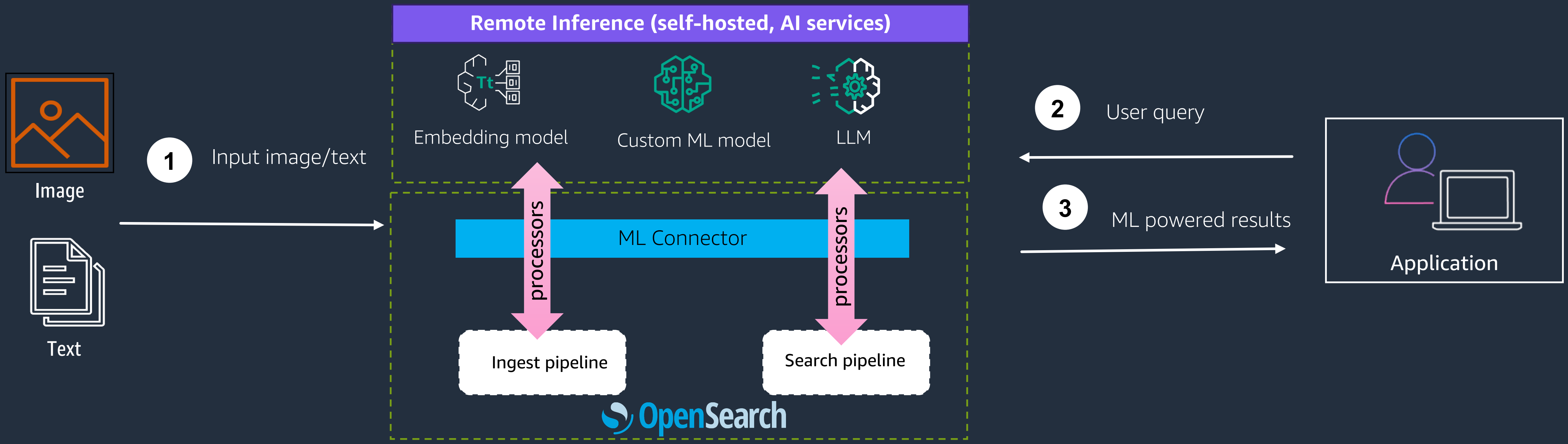
Top 4 search engine
DB-Engines ranking

75+
Partners and growing

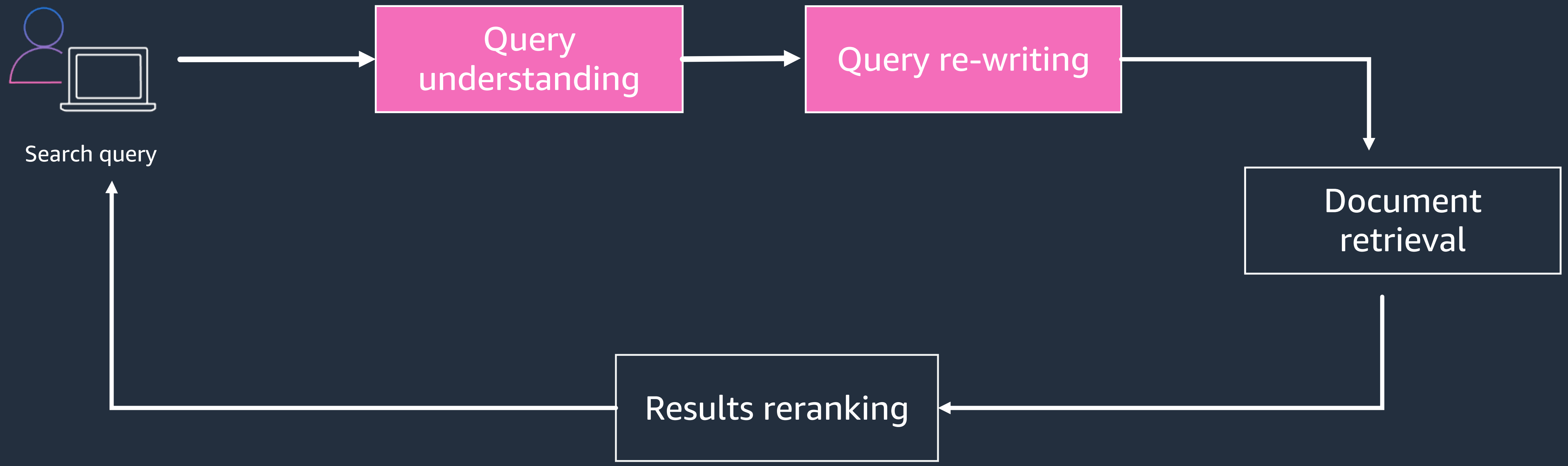
100s of new features
32k+ pull requests merged
100+ weekly community contributions

Multiple service providers
AWS, Oracle, Aiven – Azure and GCP, Bonsai-Azure and GCP

ML integrations made easy with OpenSearch



ML-Powered Search Lifecycle



Query understanding starts at ingestion time

Amazon Rekognition
or
Object detection models

```
{

"description": "Rugged Brown Leather Boots”,
“price”: “50
"color":"brown",
"category":"Apparel and Accessories",
"objects":"Footwear,Boot,Shoe,Clothing”

}
```

```
{
"Text": "Bob ordered two
sandwiches and three ice cream
cones today from a store in
Seattle.”
}
```

Amazon Comprehend
or
NER models

```
{

 {"Text": "Bob", "Type": "PERSON" },
 {"Text": "two sandwiches", "Type": "QUANTITY" },
 {"Text": "three ice cream cones", "Type": "QUANTITY" },
 {"Text": "today", "Type": "DATE" },
 {"Text": "Seattle", "Type": "LOCATION" }

}
```

Ingest pipeline: ML inference Processor

And applies at search time

User Query

“Brown leather shoes for men under 50$”

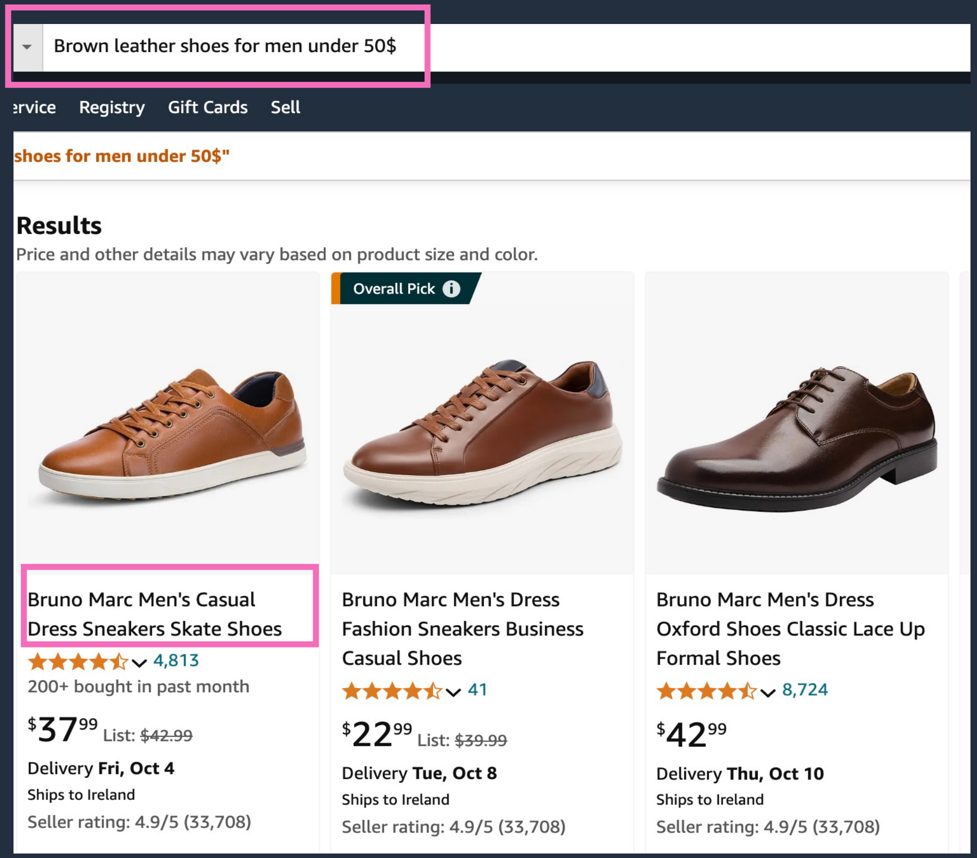
Search Query

```
{
  "query": {
    "bool": {
      "must": [
        { "match":{ "description":"shoes"}}
              ],

      "filter": [
        { "term":   { "category": "footwear" }},
        { "range": { "price":{ "lt": 50 }},
        { "term":   { "gender":"male" }},
        { "term":   { "color": "brown"}}
      ]
    }
  }
}
```

“nearby” => “term”: {“location”: “Berlin”}

“Now” => “term”: {“time”: “now”}



Brown leather shoes for men under 50$

ervice Registry Gift Cards Sell

shoes for men under 50$"

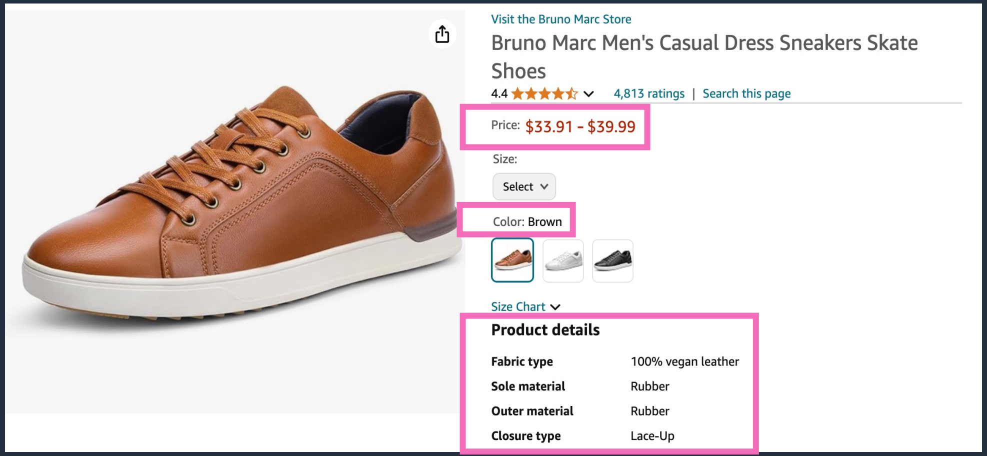
Results

Price and other details may vary based on product size and color.

Bruno Marc Men's Casual Dress Sneakers Skate Shoes
4,813
200+ bought in past month
$37.99 List: ~~$42.99~~
Delivery **Fri, Oct 4**
Ships to Ireland
Seller rating: 4.9/5 (33,708)

Overall Pick
Bruno Marc Men's Dress Fashion Sneakers Business Casual Shoes
41
$22.99 List: ~~$39.99~~
Delivery **Tue, Oct 8**
Ships to Ireland
Seller rating: 4.9/5 (33,708)

Bruno Marc Men's Dress Oxford Shoes Classic Lace Up Formal Shoes
8,724
$42.99
Delivery **Thu, Oct 10**
Ships to Ireland
Seller rating: 4.9/5 (33,708)



Visit the Bruno Marc Store

Bruno Marc Men's Casual Dress Sneakers Skate Shoes

4.4 | 4,813 ratings | Search this page

Price: $33.91 - $39.99

Size: Select

Color: Brown

Size Chart

Product details

Fabric type	100% vegan leather
Sole material	Rubber
Outer material	Rubber
Closure type	Lace-Up

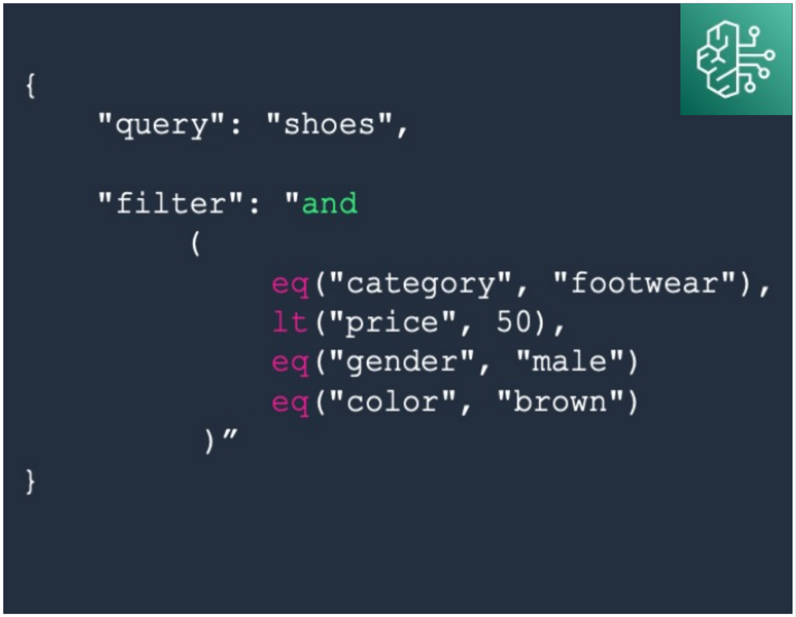
Query re-writing using LLMs

Search Query

“Brown leather shoes for men under 50$”

→

Structured Query

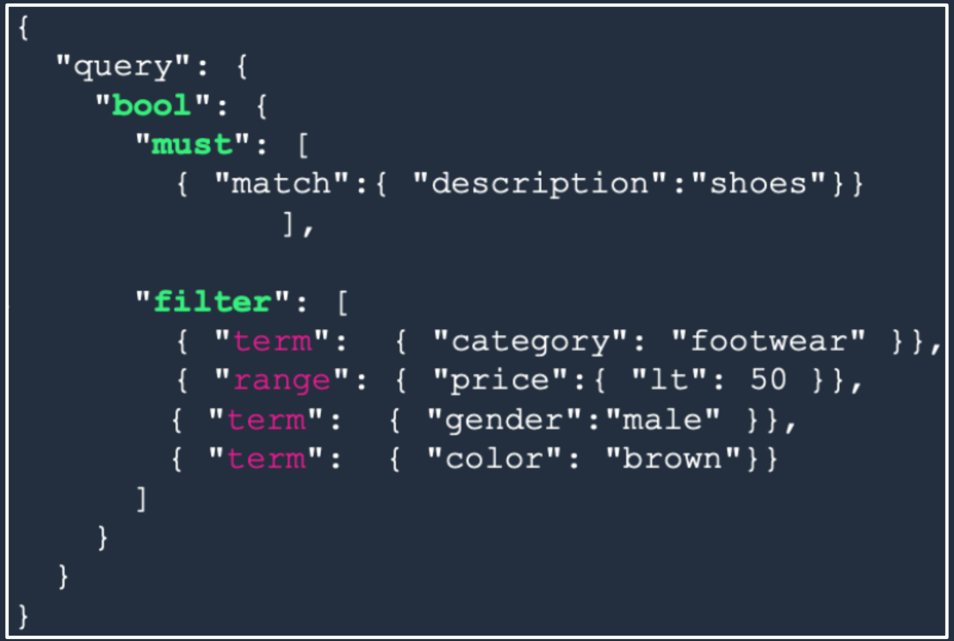


```
{
    "query": "shoes",

    "filter": "and
          (
              eq("category", "footwear"),
              lt("price", 50),
              eq("gender", "male")
              eq("color", "brown")
          )"
}
```

Rewrite →

OpenSearch query DSL

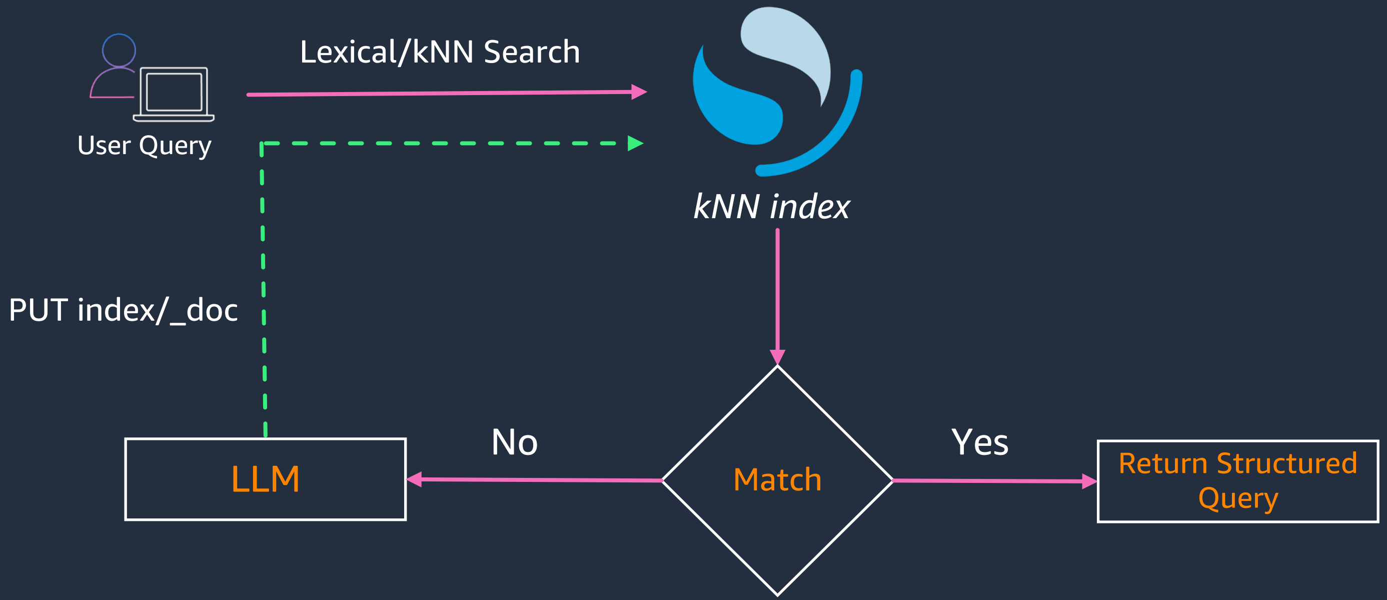


```
{
  "query": {
    "bool": {
      "must": [
        { "match":{ "description":"shoes"}}
            ],

      "filter": [
        { "term":   { "category": "footwear" }},
        { "range": { "price":{ "lt": 50 }},
        { "term":   { "gender":"male" }},
        { "term":   { "color": "brown"}}
      ]
    }
  }
}
```

GitHub code sample

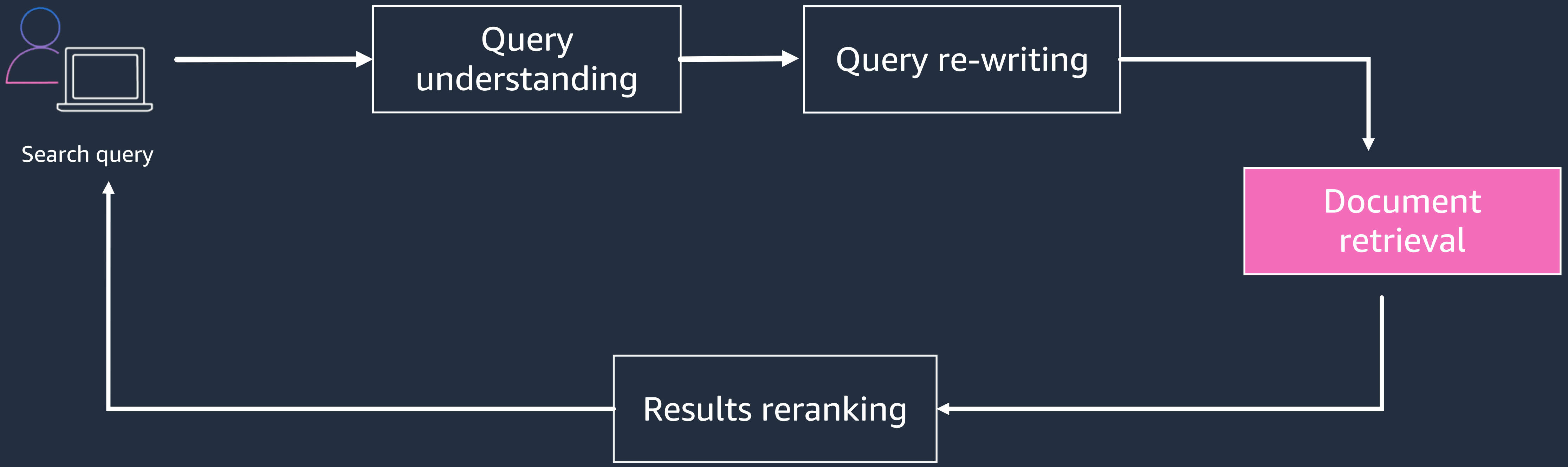
Cache to reduce the LLM cost and latency



Search latency
3 sec → 600 ms

- Warm up cache on regular intervals
- ISM for auto-deleting the cache (index) based on TTL
- Langchain built-in semantic cache with OpenSearch

ML-Powered Search Lifecycle



Retrieval Types

Sparse retrieval

BM25

Lexical search

Uses keyword-based matching using TF/IDF.

Neural sparse search

Uses `amazon/neural-sparse` embedding models

Dense retrieval

KNN/ANN

Vector search

Uses purpose build models for text, image or video embedding models.

Multimodal search

Uses embedding models that share embedding space between image and text.

Hybrid Search

BM25+KNN

Ensemble search

Uses both traditional keyword-search combined with vector search

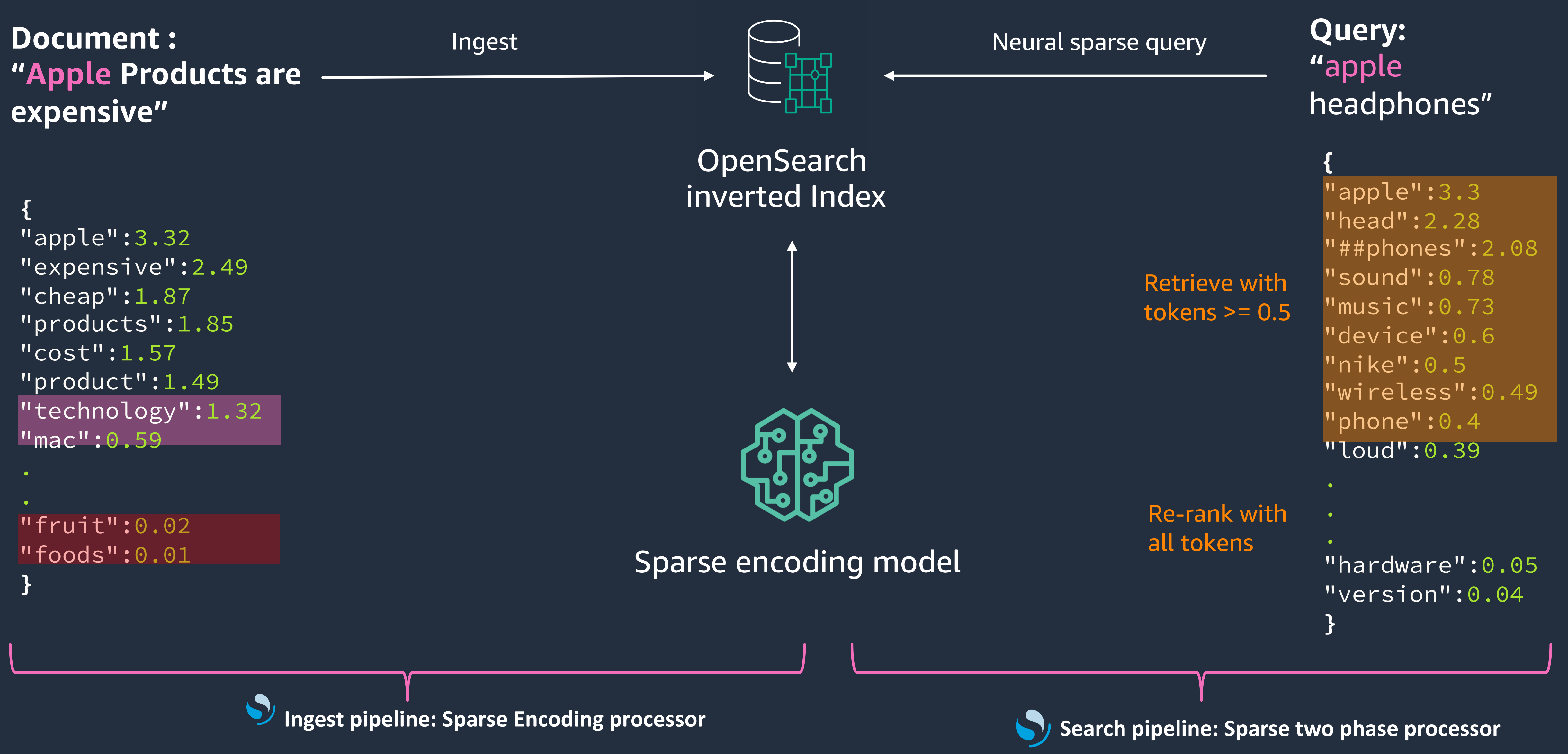
Conversational search

(BM25/KNN)+LLM

RAG

Uses LLM to augment the results retrieved from Vector Search.

Neural sparse search



Document :
“Apple Products are expensive”

Ingest

OpenSearch inverted Index

Neural sparse query

Query:
“apple headphones”

```
{
"apple":3.32
"expensive":2.49
"cheap":1.87
"products":1.85
"cost":1.57
"product":1.49
"technology":1.32
"mac":0.59
.
.
"fruit":0.02
"foods":0.01
}
```

Sparse encoding model

Retrieve with tokens >= 0.5

Re-rank with all tokens

```
{
"apple":3.3
"head":2.28
"##phones":2.08
"sound":0.78
"music":0.73
"device":0.6
"nike":0.5
"wireless":0.49
"phone":0.4
"loud":0.39
.
.
.
"hardware":0.05
"version":0.04
}
```

Ingest pipeline: Sparse Encoding processor

Search pipeline: Sparse two phase processor

Improved sparse search performance maintaining relevance

OpenSearch Sparse models v2

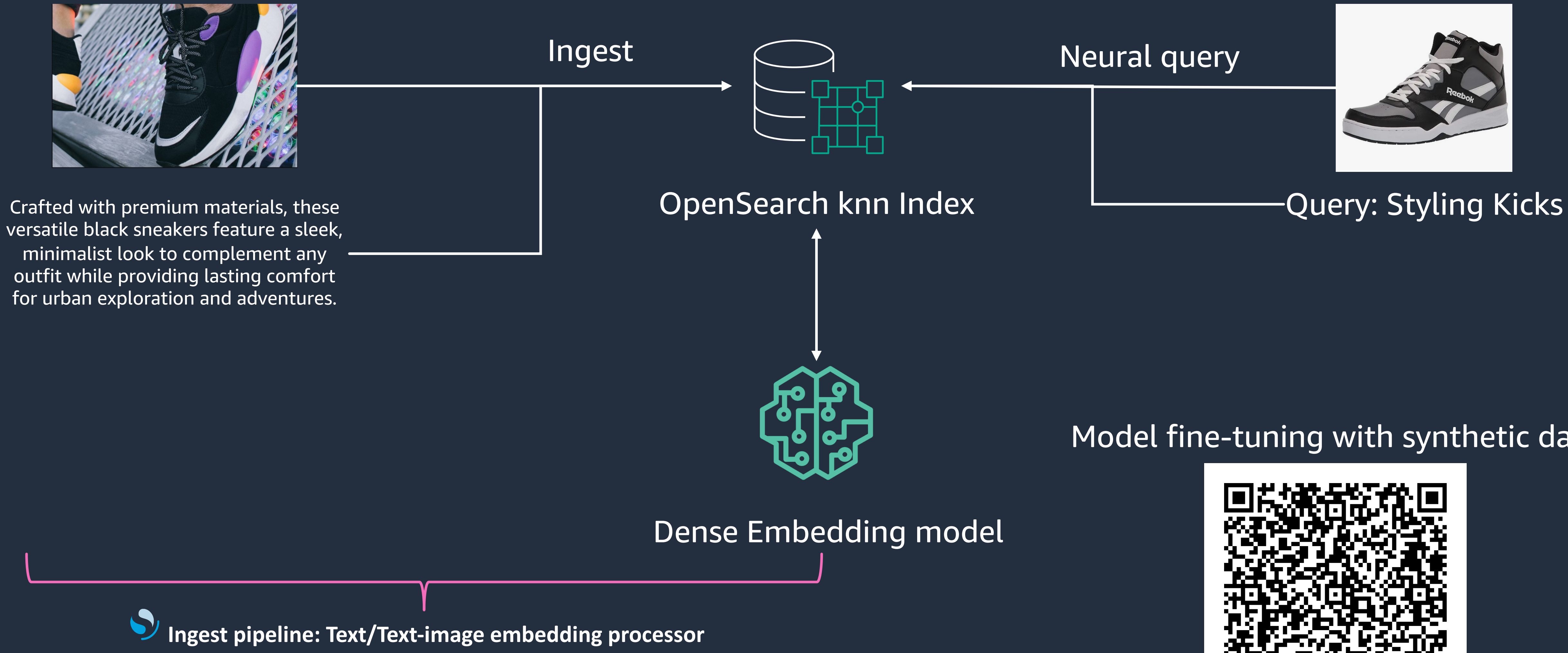
1. opensearch-neural-sparse-encoding-v2-distill
2. opensearch-neural-sparse-encoding-doc-v2-distill
3. opensearch-neural-sparse-encoding-v2-mini

https://huggingface.co/opensearch-project

Performance @ CPU	VS Sparse models v1
Ingestion ThroughPut	↑ 1.74x – 4.18x
Search Latency	↓ 30%
BIER: ndcg@10	~=

https://opensearch.org/blog/neural-sparse-v2-models

Dense search



Ingest

Neural query

Crafted with premium materials, these versatile black sneakers feature a sleek, minimalist look to complement any outfit while providing lasting comfort for urban exploration and adventures.

OpenSearch knn Index

Query: Styling Kicks

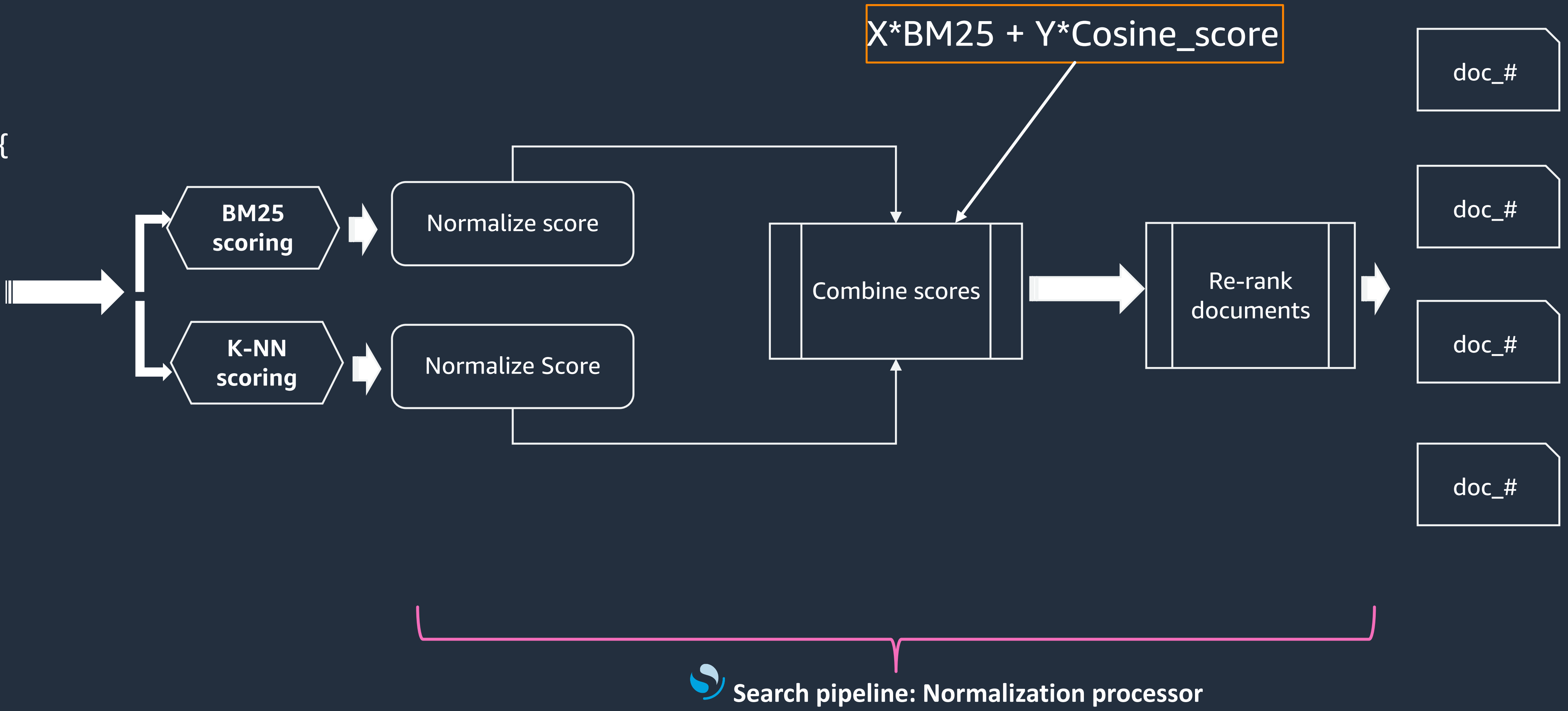
Dense Embedding model

Ingest pipeline: Text/Text-image embedding processor

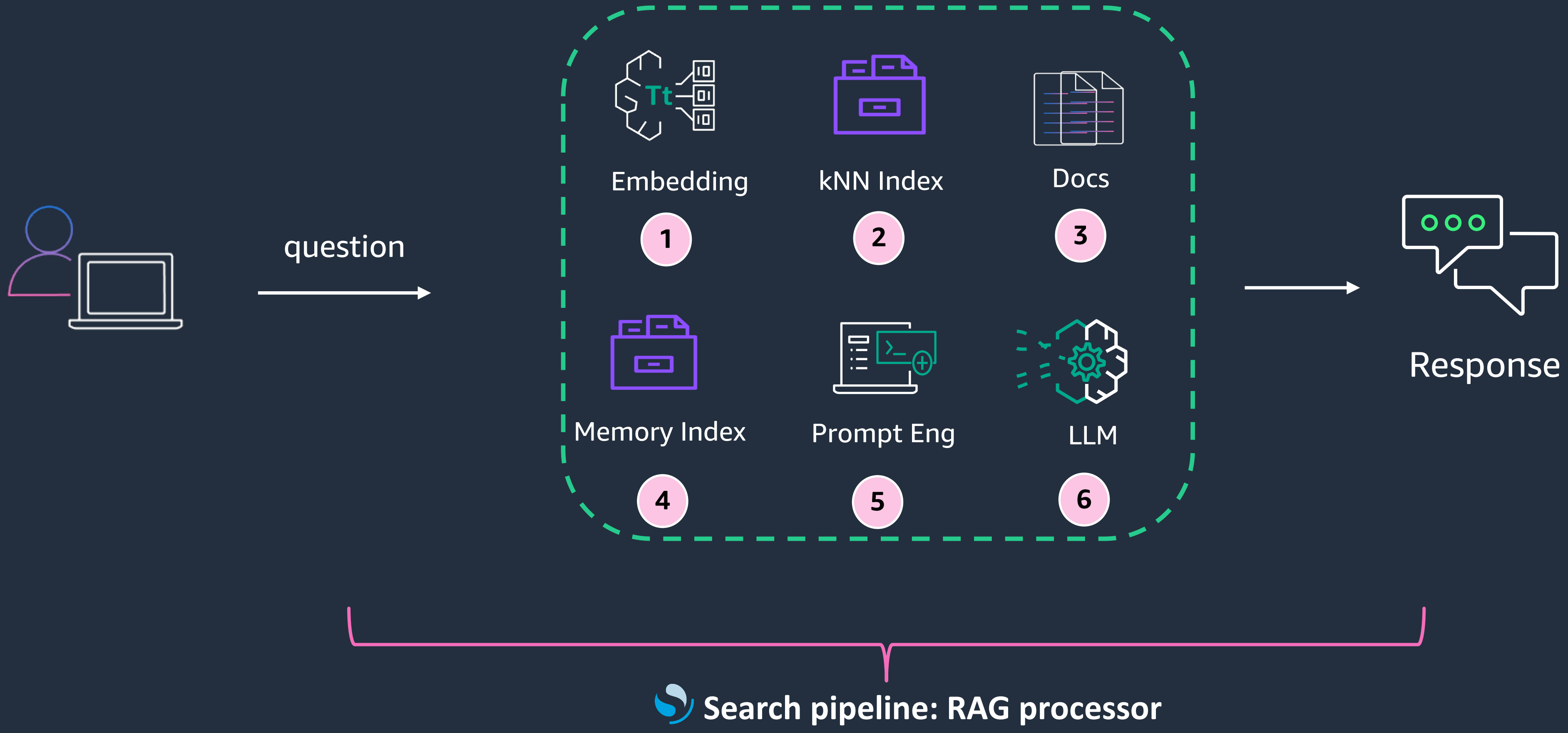
Model fine-tuning with synthetic data

Hybrid search

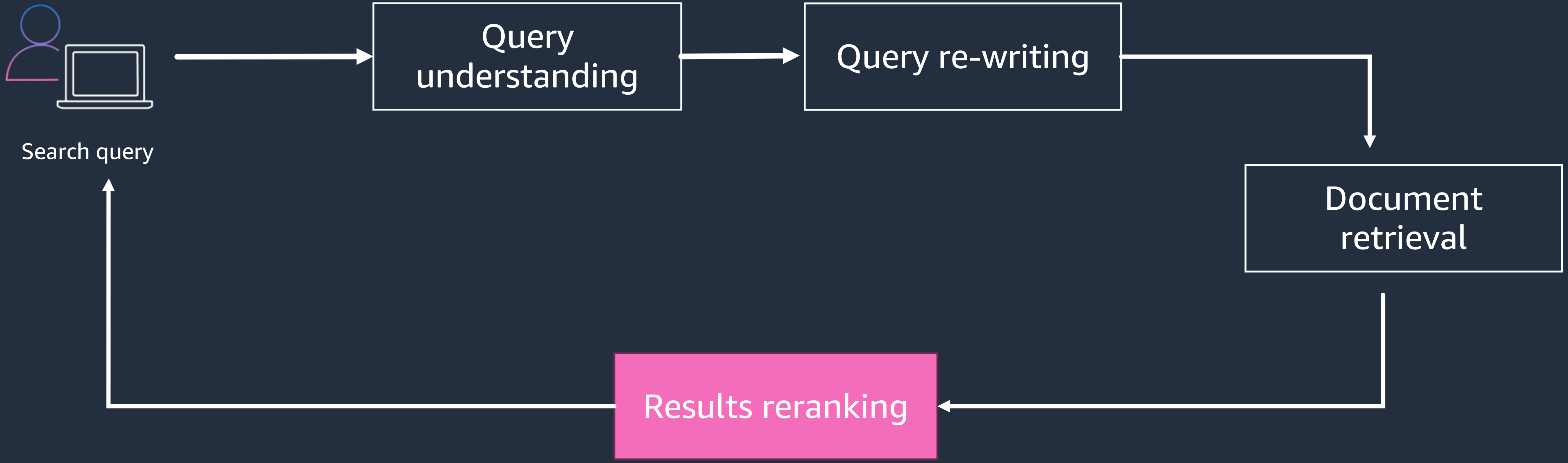
```
"query": { "hybrid": {
"queries": [ {
    “match”:{
        <text query>
    }
},
  {
    “neural”:{
       <vector query>
    }
}]
```



Conversational search



ML-Powered Search Lifecycle



Search query

Query understanding

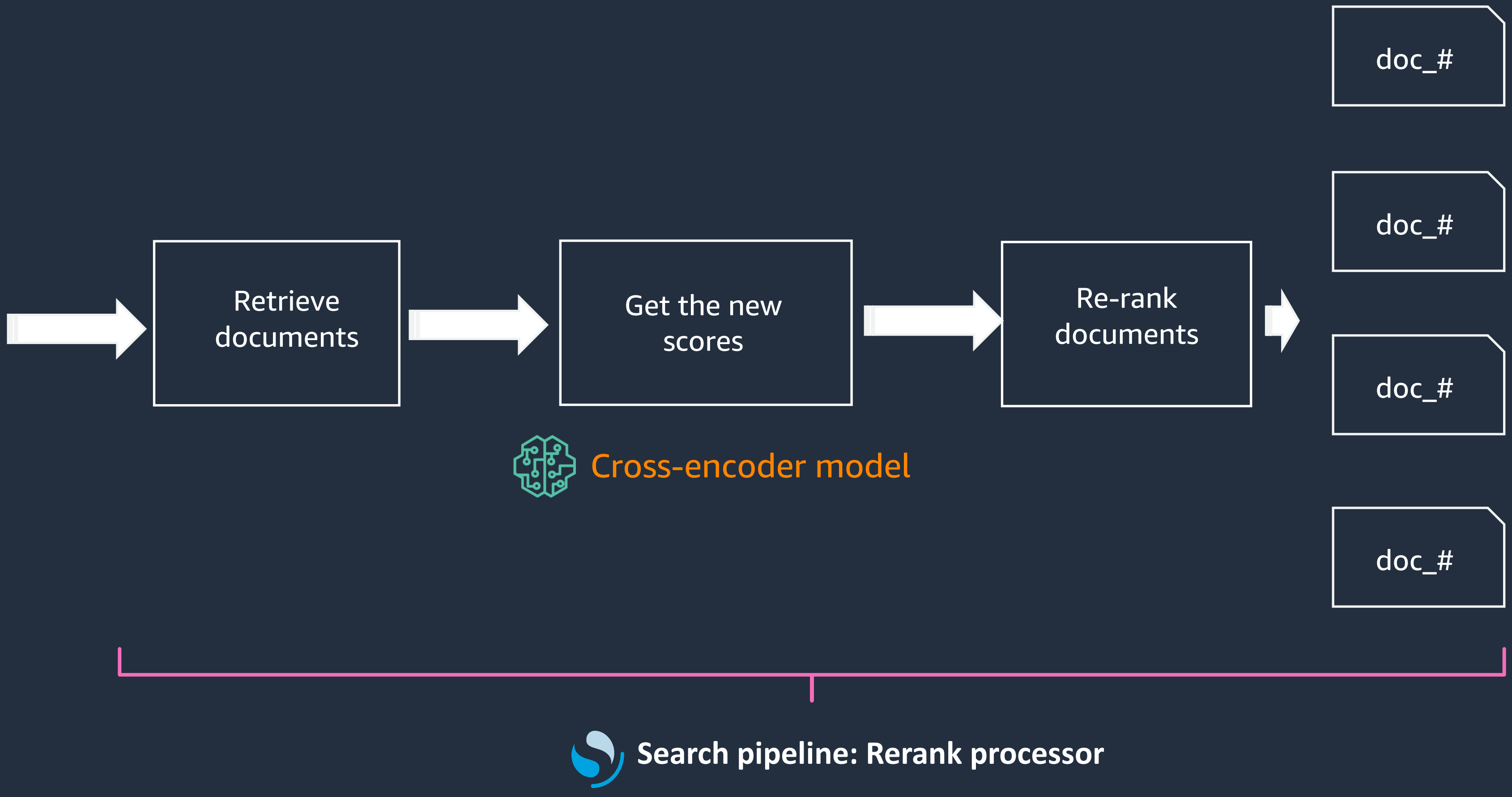
Query re-writing

Document retrieval

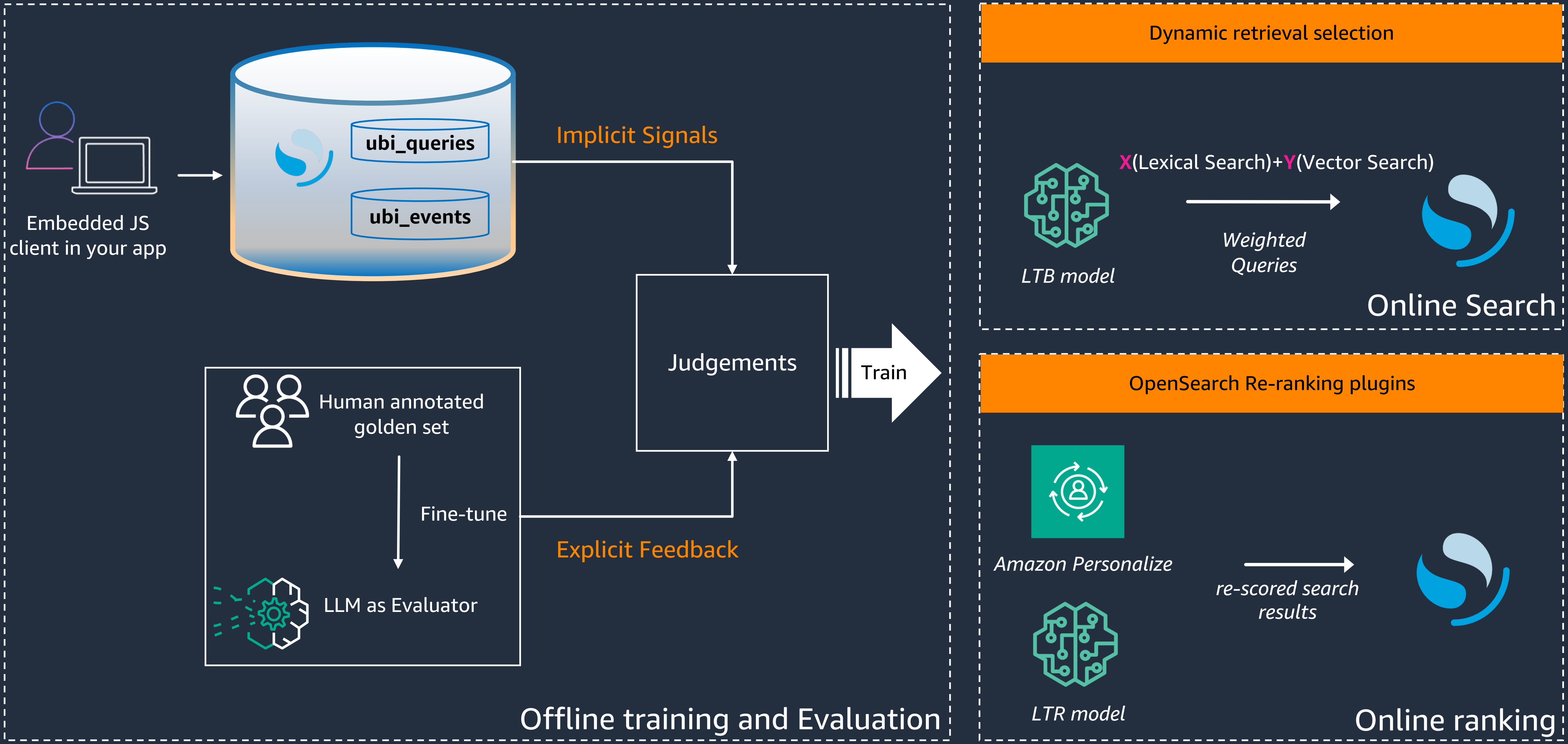
Results reranking

Reranking with cross-encoder models

```
{
 "query": {
  "match": {
   "text_representation": ""…
  }
 },
 "ext": {
  "rerank": {
   "query_context": {
    "query_text": "Where is Albuquerque?"
   }
  }
 }
}
```



User Behavior insights to improve your search and ranking



Demo

Thank you!

Hajer Bouafif

in linkedin.com/in/hajerbouafif

Praveen Mohan

in linkedin.com/in/praveen-mohanprasad