WORKSHOP

Generative AI

Valentina Alto

Azure Specialist —
Data and Artificial Intelligence
Microsoft

Multimodal Retrieval Augmented Generation

VALENTINA ALTO

APRIL 23, 2024

Getting Started with MM RAG

An implementation with GPT-4v and Llama-Index

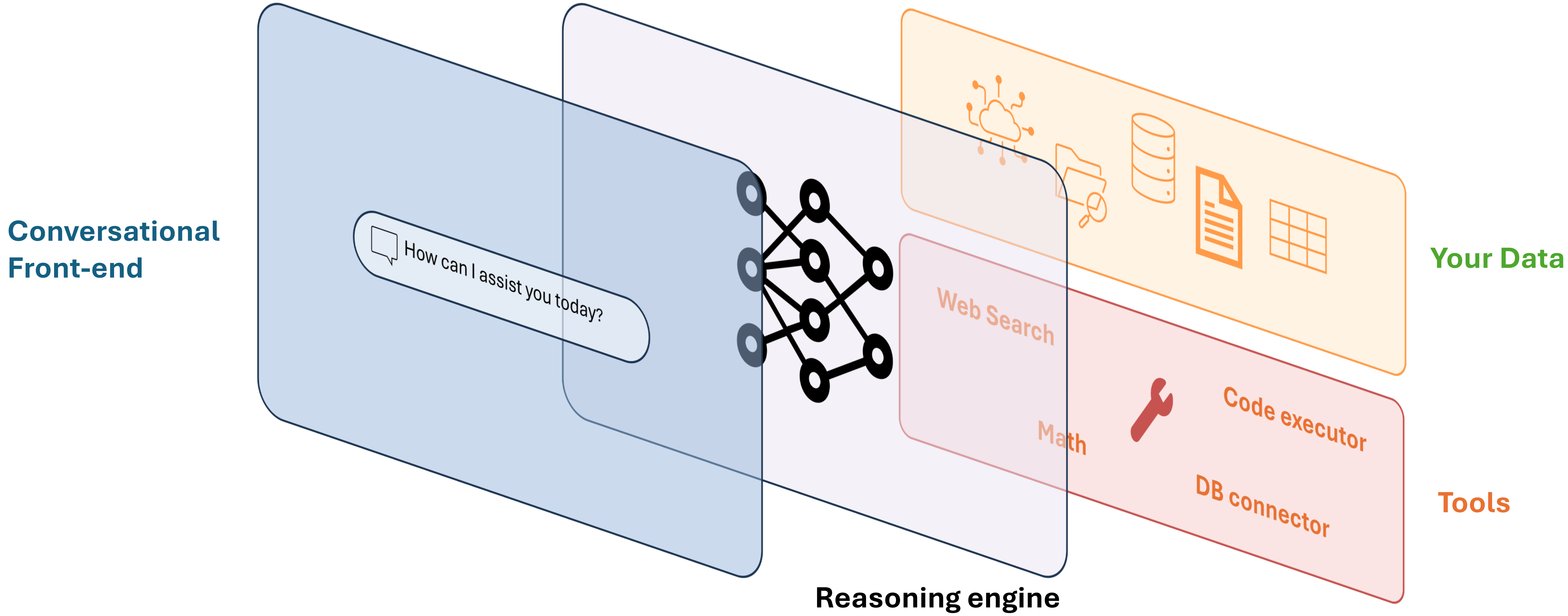
Agenda

1. Intro to RAG
2. Embeddings and VectorDB
3. Multimodal RAG
4. Demo Time

Tip: Download the workshop material to start familiarizing with code and slides.

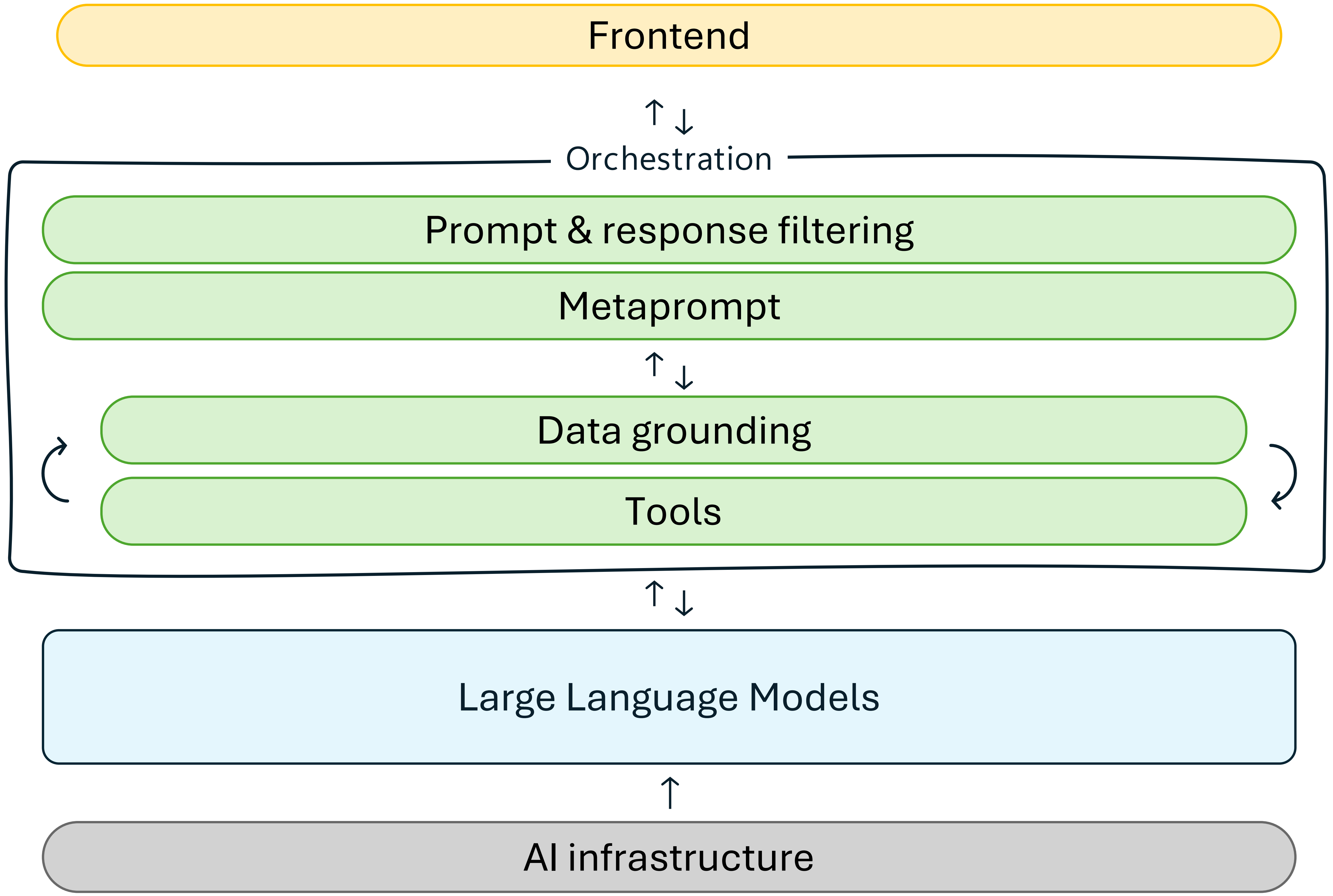
How: Scan the QR code below!

LLMs as "brains" for our applications



Anatomy of an LLM-powered application

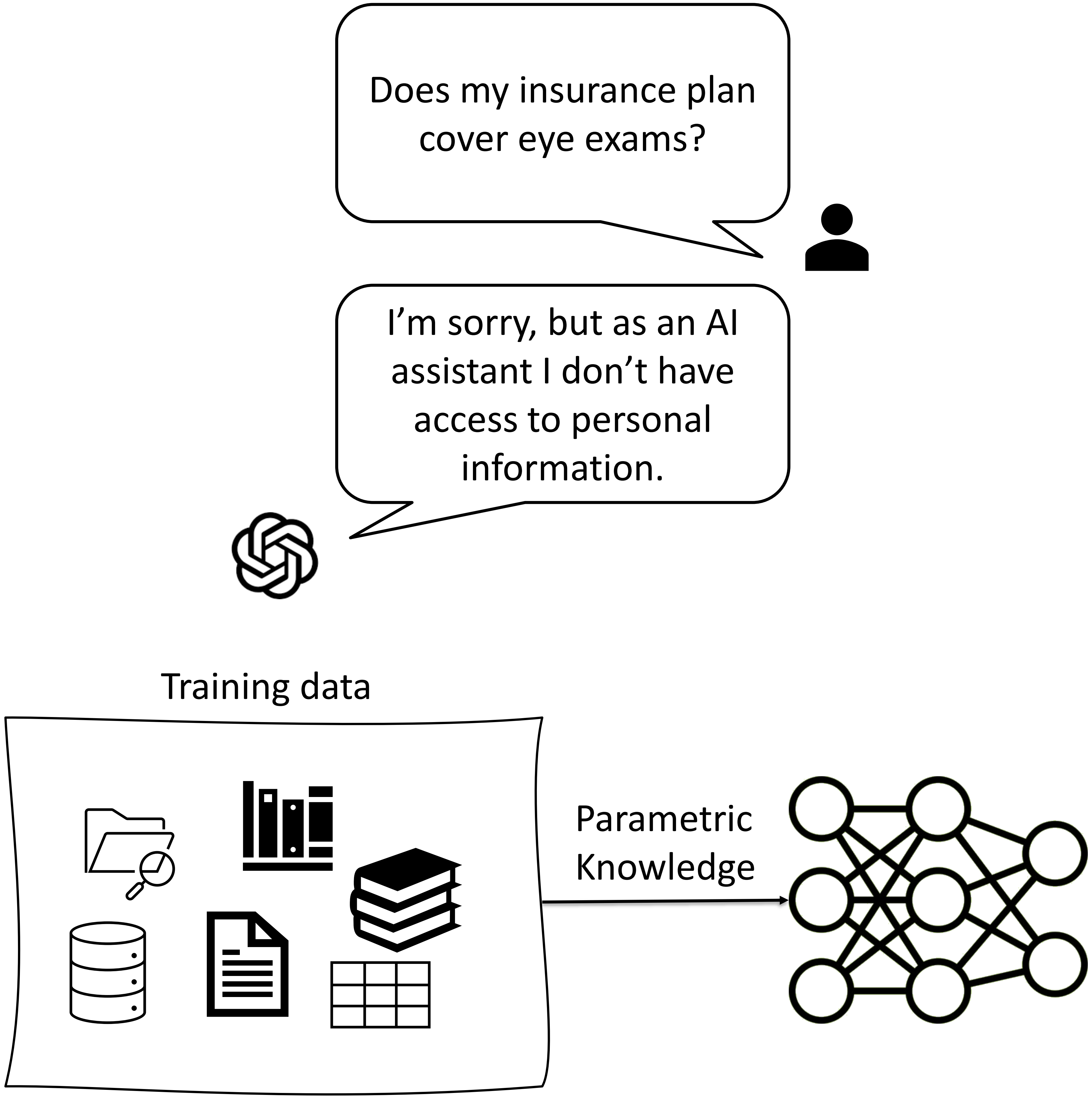
LLM-powered applications open the way to a new landscape of components



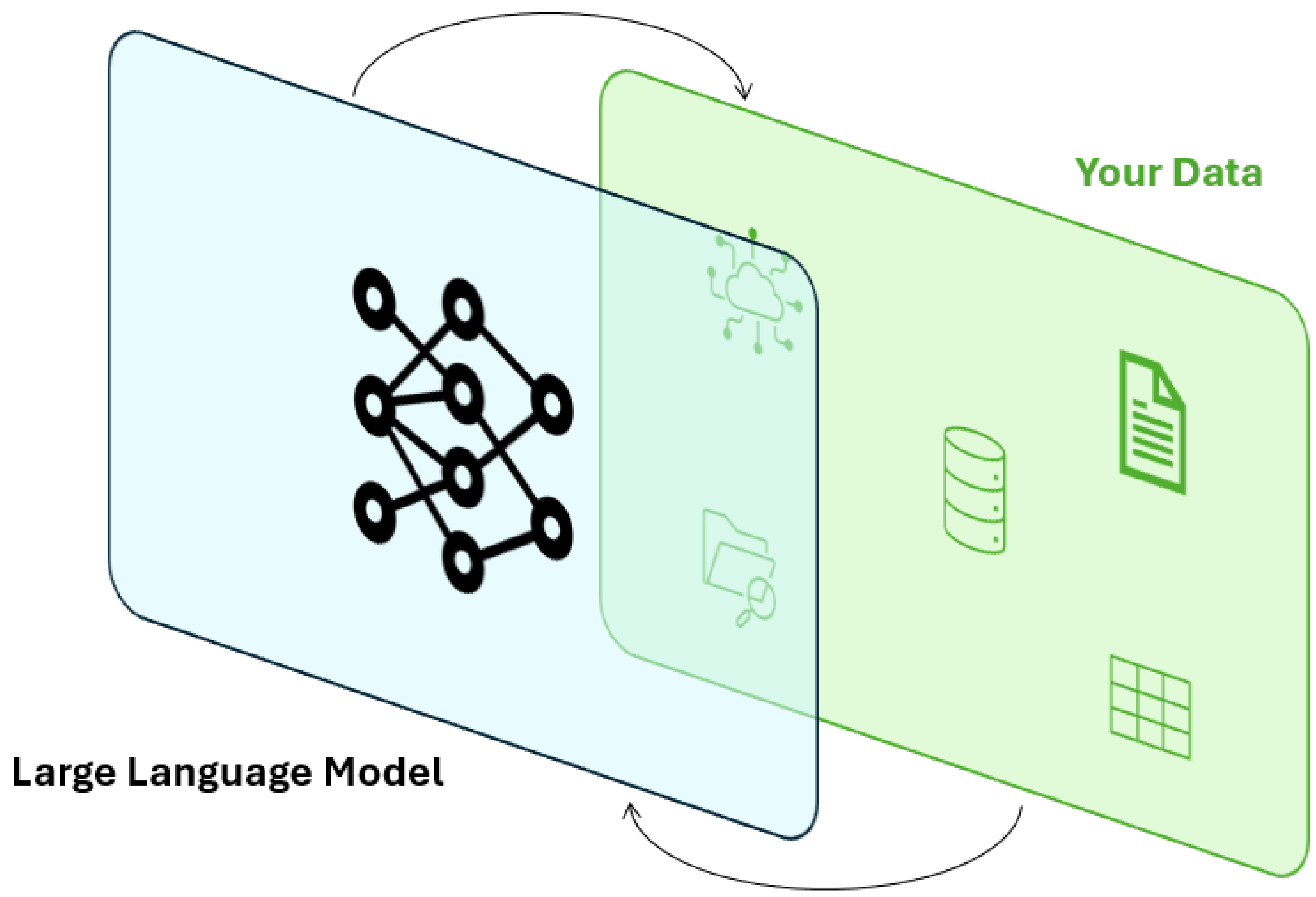
Problem: Generative AI doesn't know about your data

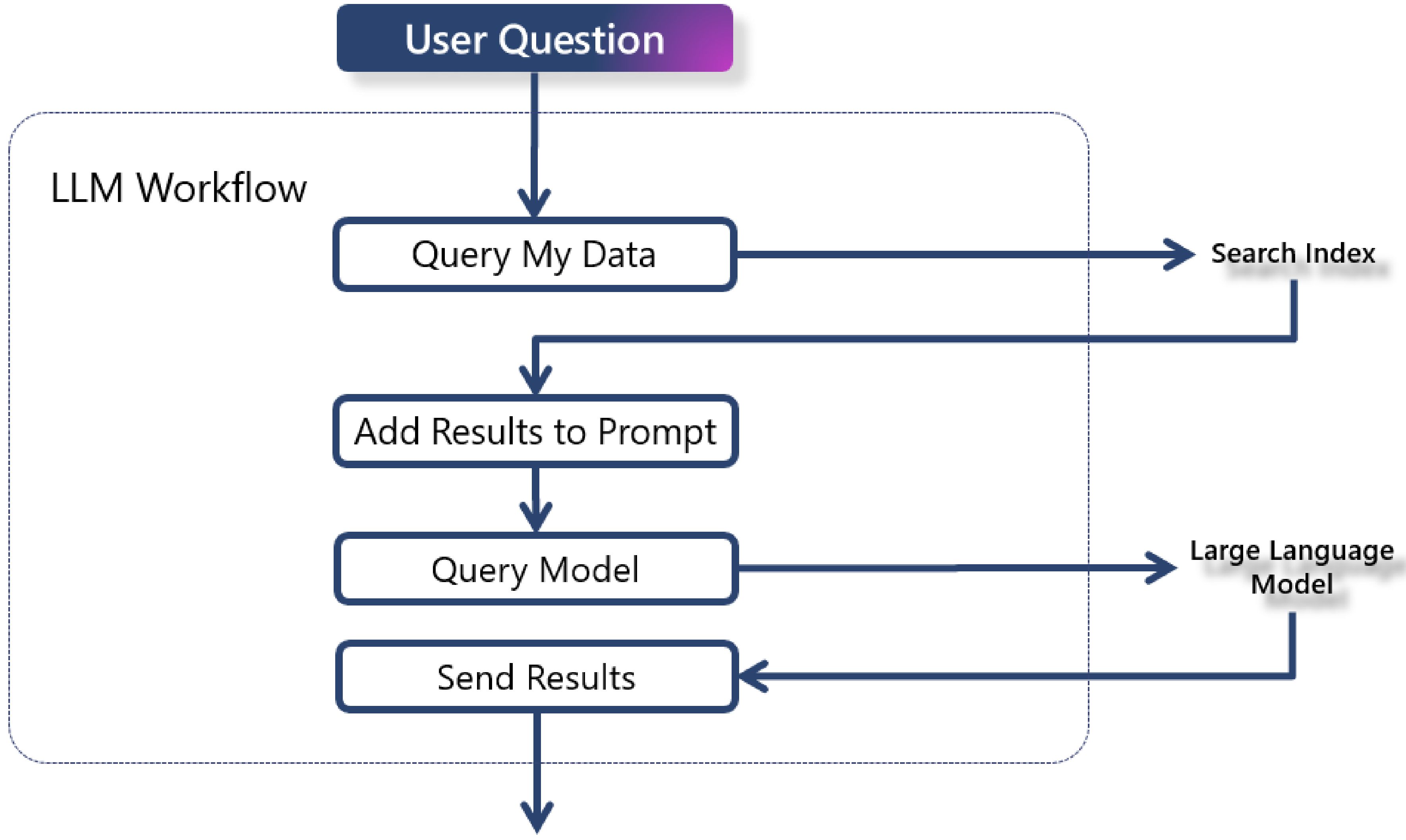
What if data you are interested in are not part of the training dataset?

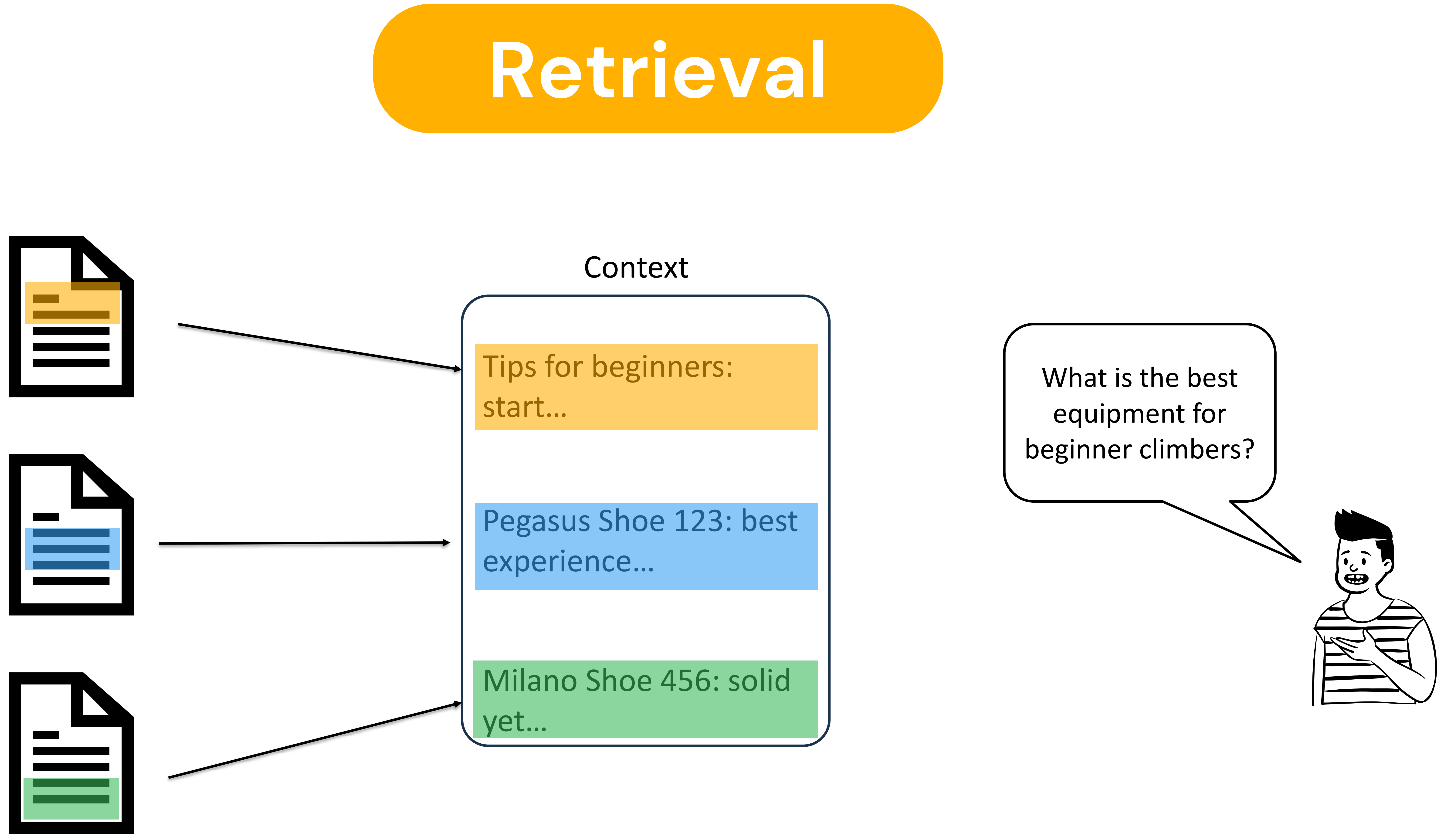
- Personal Data (confidential, not public…)
- Up to date data
- Application data



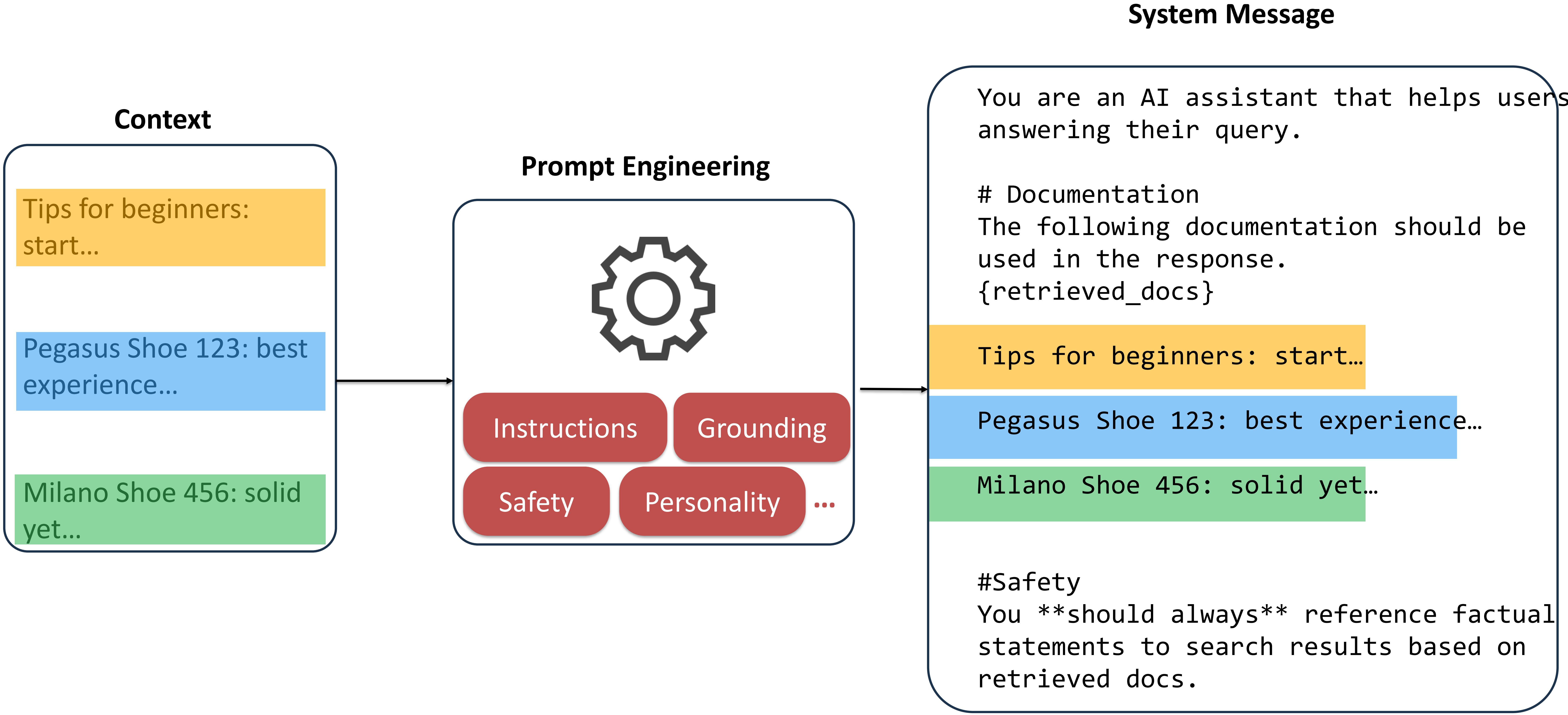
Introducing Retrieval Augmented Generation






Retrieval
Context
Tips for beginners: start...
Pegasus Shoe 123: best experience...
Milano Shoe 456: solid yet...
What is the best equipment for beginner climbers?

Augmentation



Generation

System message + retrieved documents

You are an AI assistant that helps users answering their query.

Documentation
The following documentation should be used in the response.
{retrieved_docs}

Tips for beginners: start…

Pegasus Shoe 123: best experience…

Milano Shoe 456: solid yet…

#Safety
You **should always** reference factual statements to search results based on retrieved docs.

+

User's query

What is the best equipment for beginner climbers?

Generative Model
(e.g. GPT-4

"According to the catalogue, if you are about to start climbing…"

How do we retrieve relevant documents?

<3d style illustration of two cats playing a card game, one cat holding the deck of cards, the other cat thinking about which card to draw from the deck>

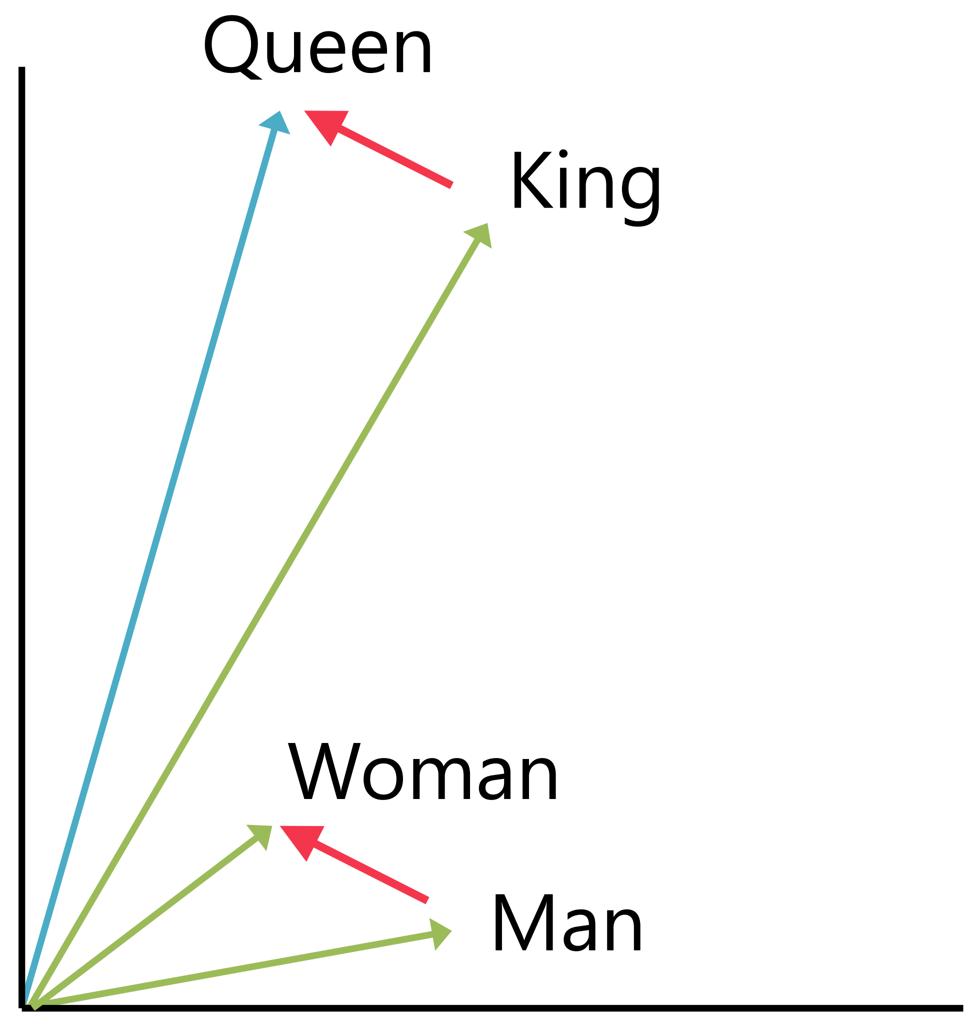
Embedding – 1/4

An embedding is a way of representing high-dimensional, non-numeric data, such as words or sentences, in a lower-dimensional space, such as vector.

A text embedding can capture the semantic and syntactic features of the text, such as meaning, context, and similarity.

Each embedding is a vector of floating-point numbers, such that the distance between two embeddings in the vector space is correlated with semantic similarity between two inputs in the original format.

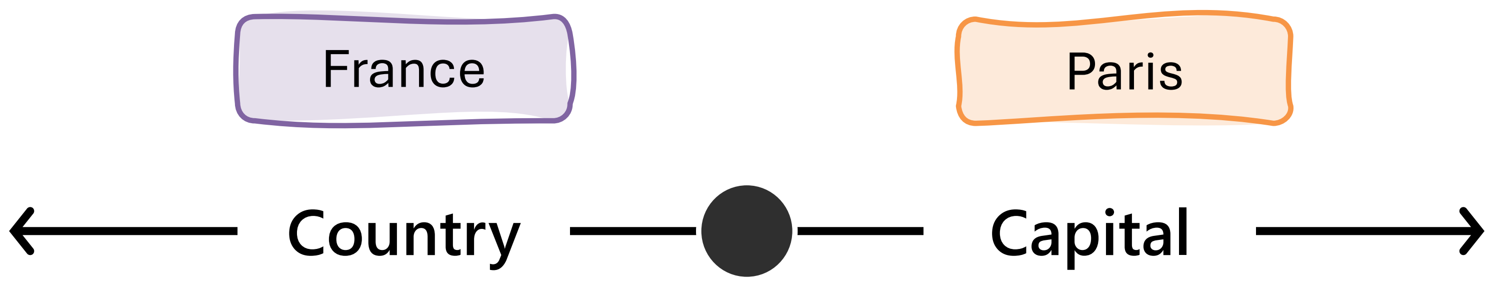
For example, if two concepts are similar, then their vector representations should also be similar.



King-Man+Woman ≈ Queen

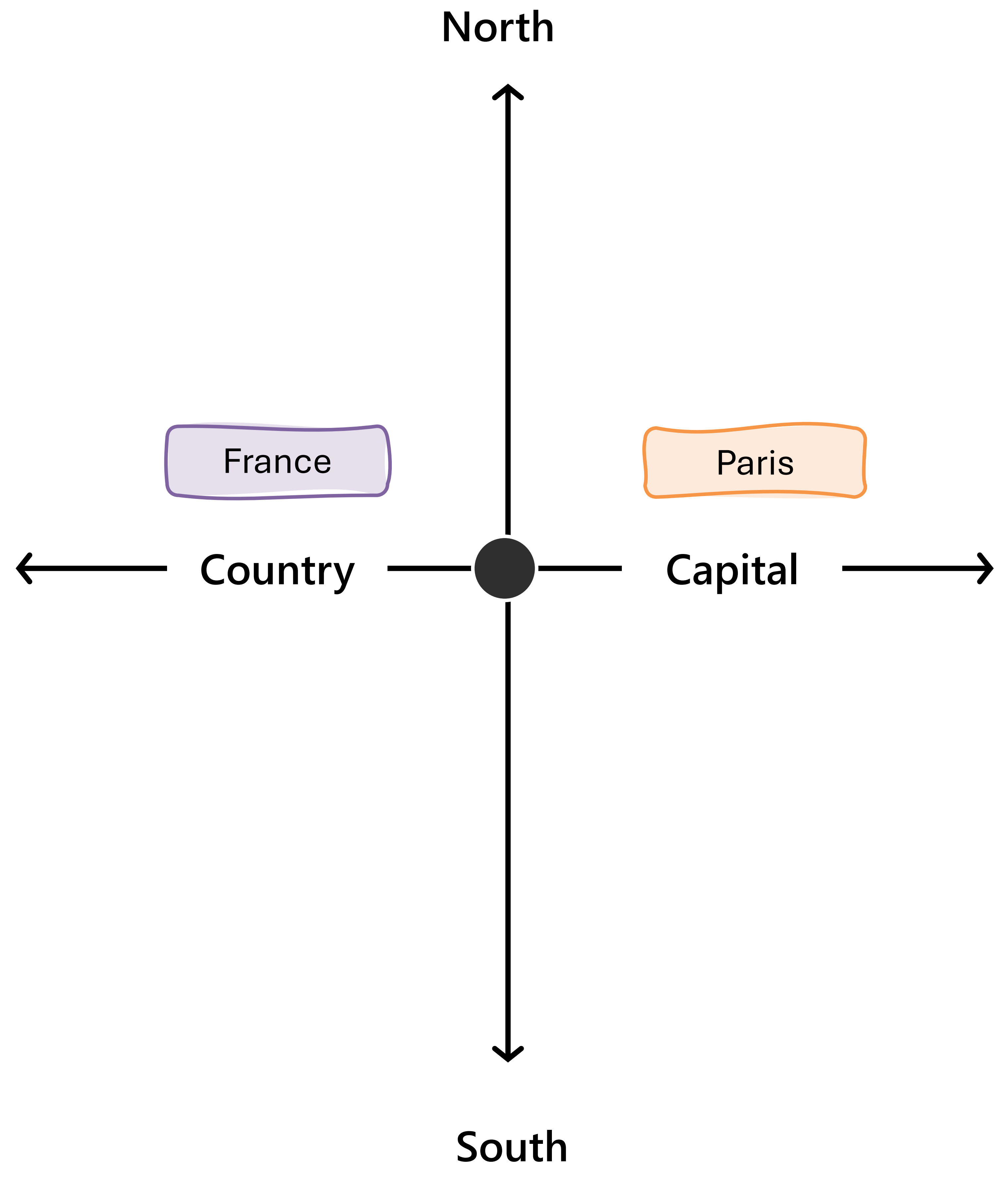
Embedding – 1/4

Embeddings represent your data, and each dimension represents a feature of that data.



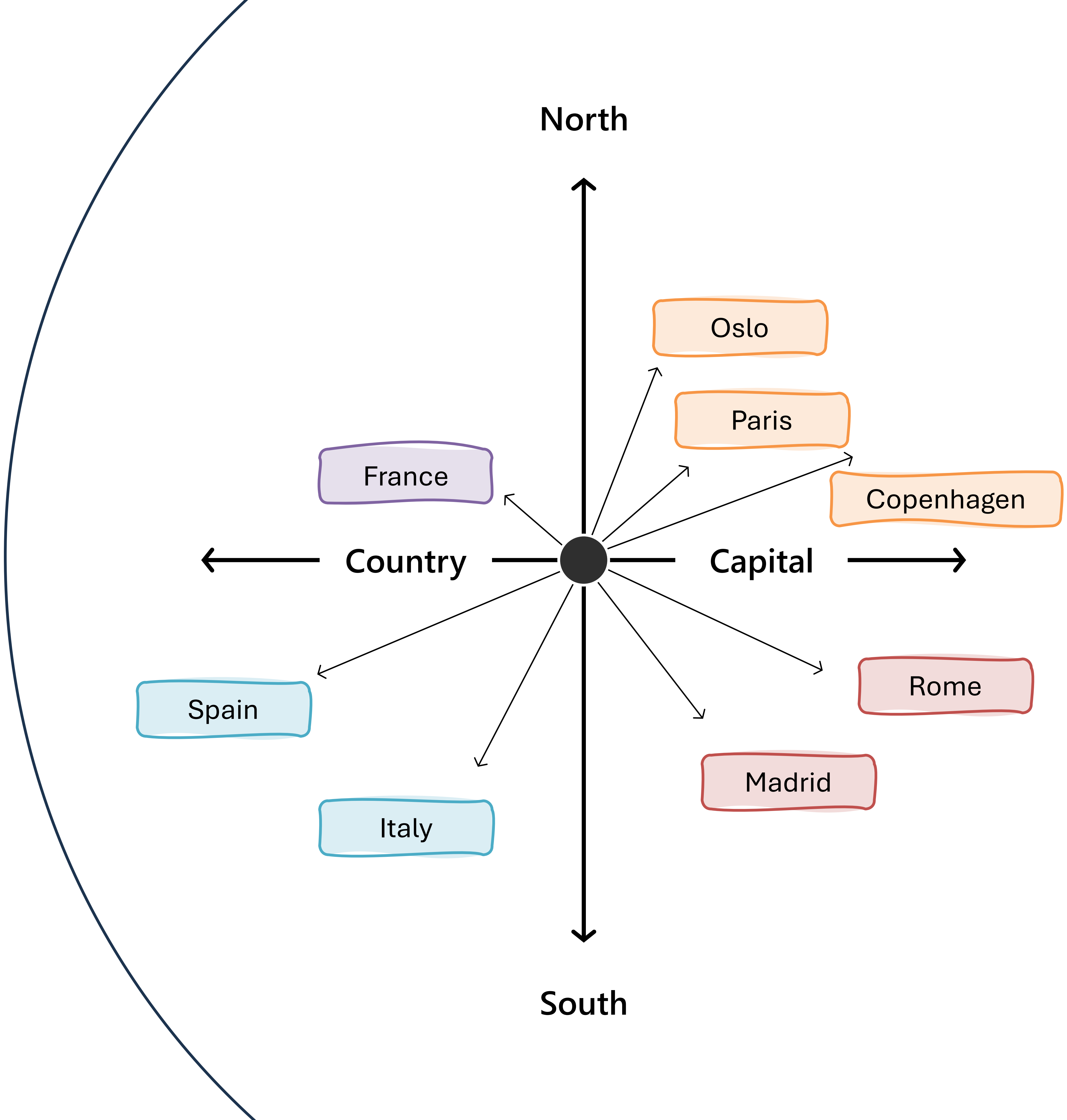
Embedding – 3/4

For example, one dimension could be the geographic connotation (country vs capital), another one the geographic position (north vs south).



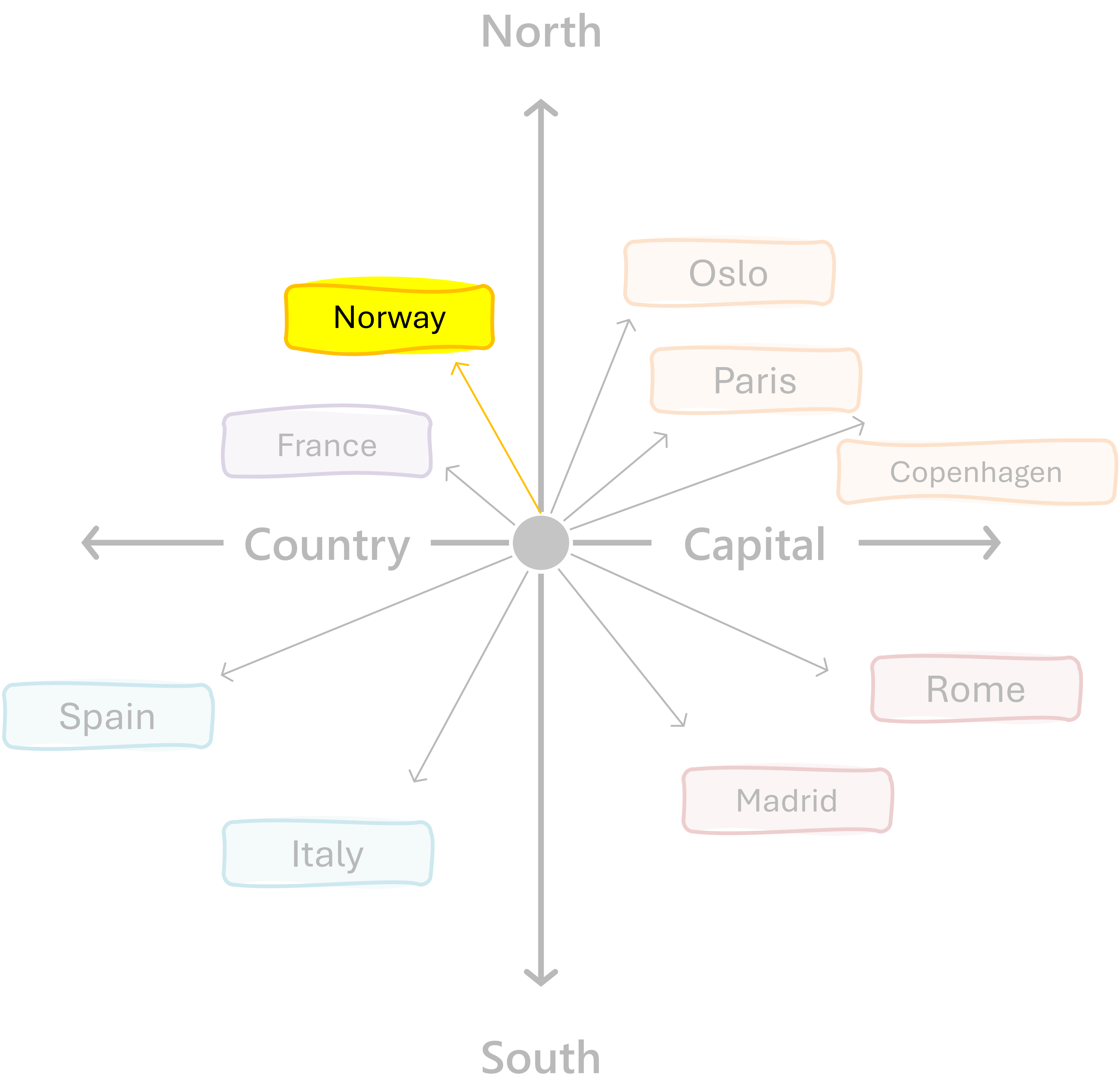
Embedding – 4/4

In the embedding space, similar concepts (words, sentences, documents) should be close in mathematical distance.



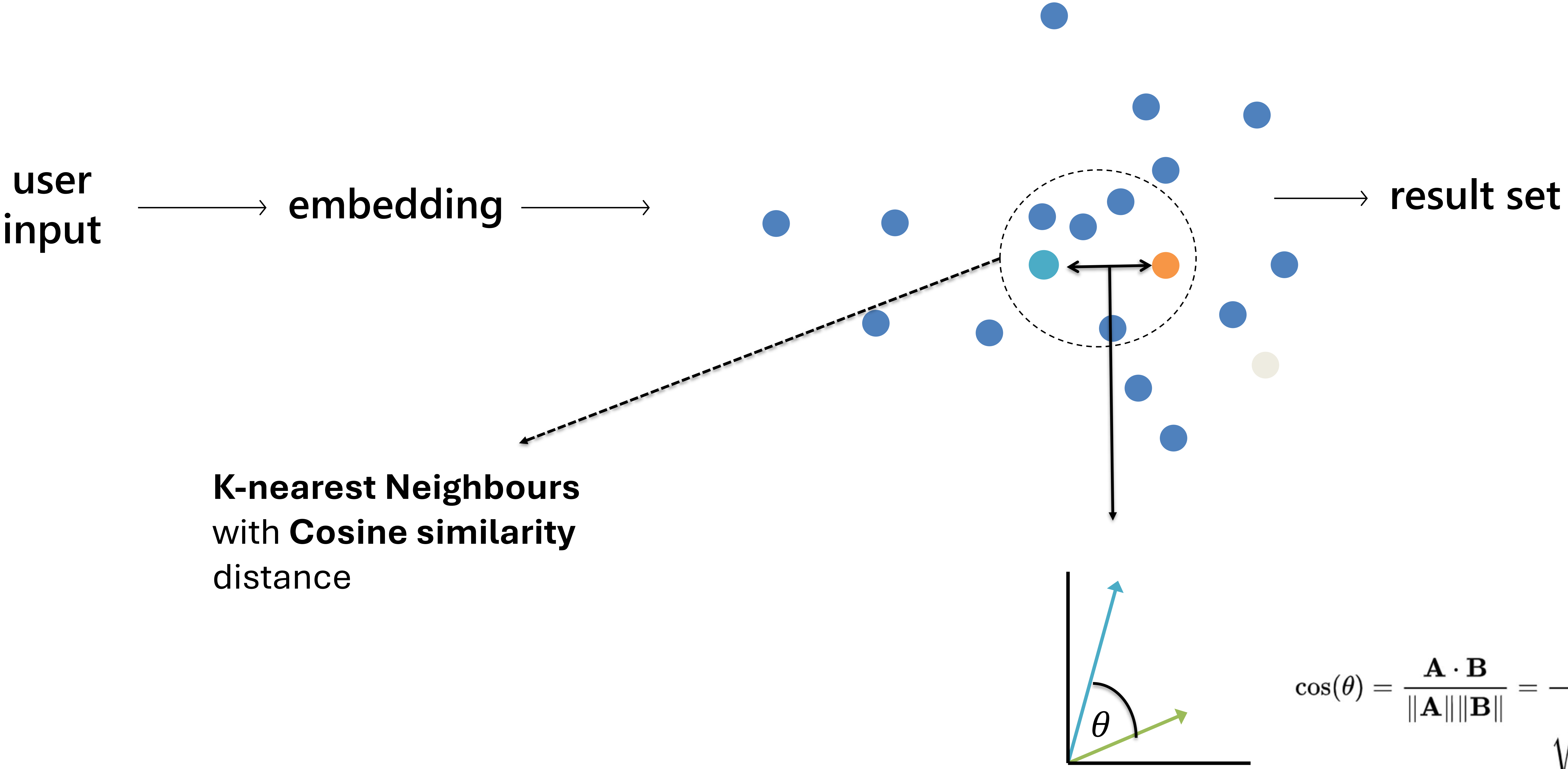
Vector search ranks objects by similarity (relevance) to the query

Norway



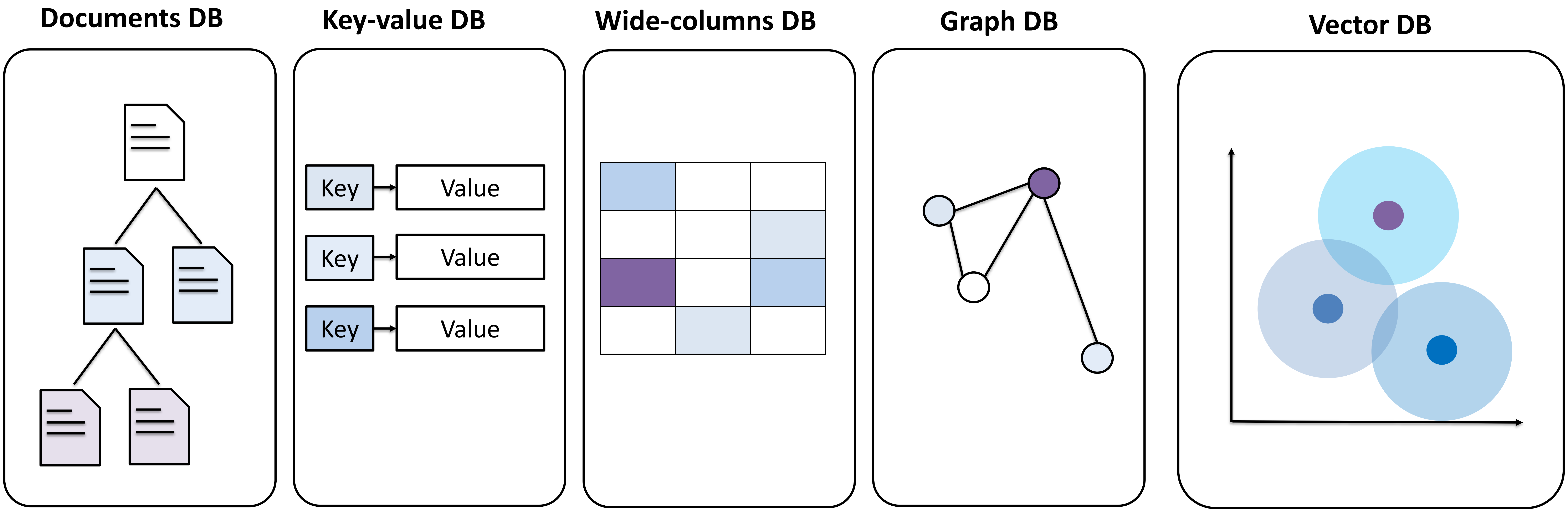
Relevance	Result
Query	Norway
1	France
2	Oslo
3	Copenhagen
4	…

Similarity Search with Embeddings

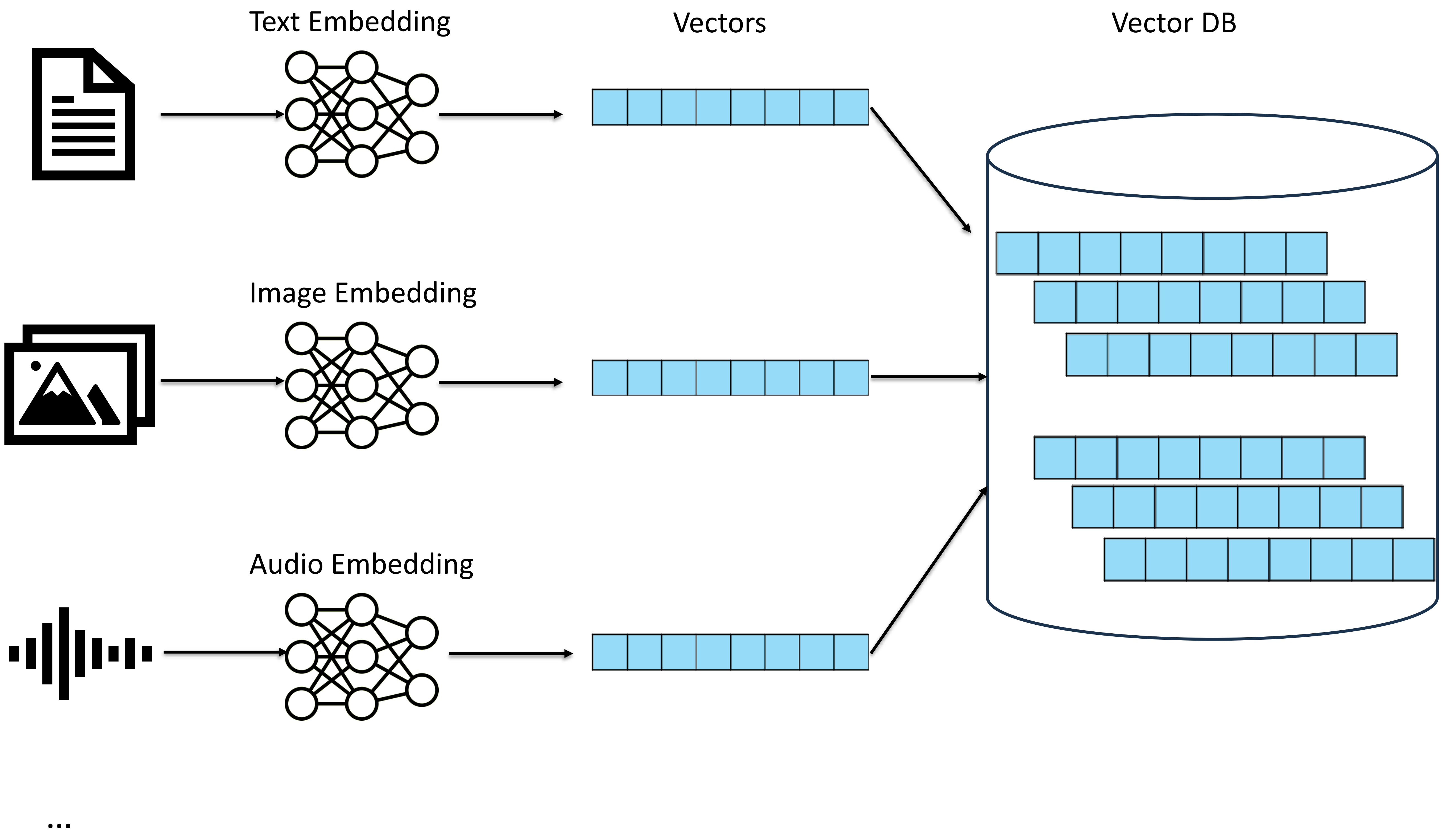


$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

A new entry in the landscape of DB



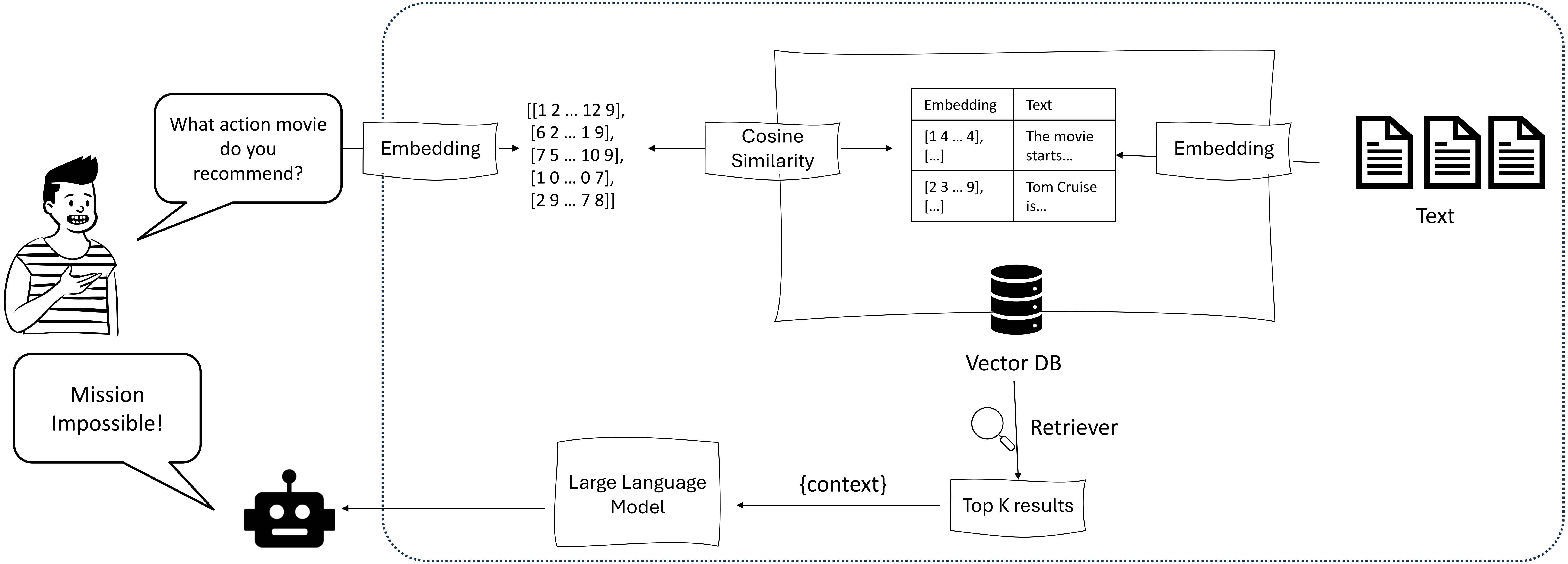
Vector Databases



chroma
marqo
vespa
qdrant
LanceDB
Milvus
Weaviate
Pinecone

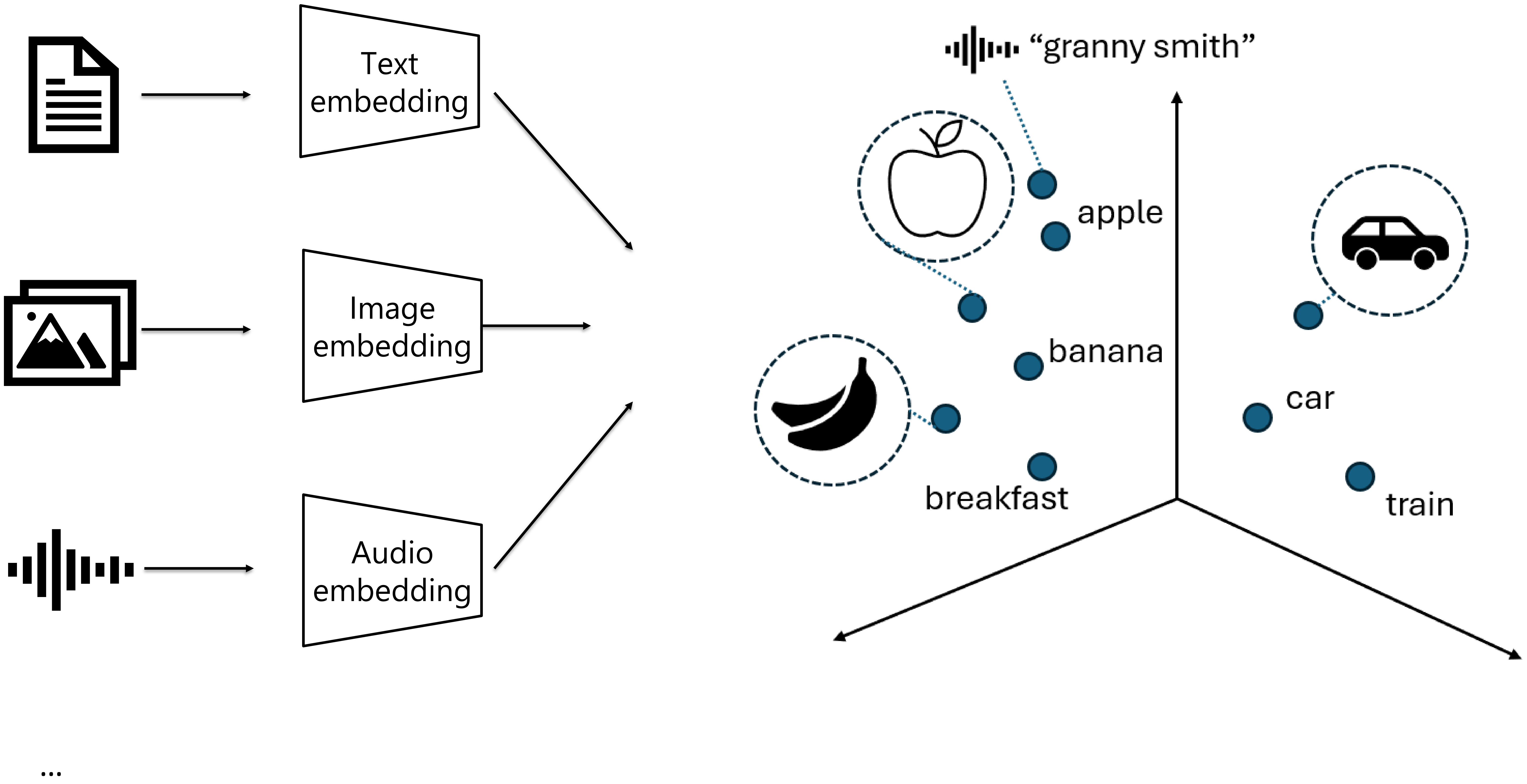
OpenSearch
ClickHouse
PostgreSQL
cassandra
elasticsearch
redis
SQL
Azure SQL

Monomodal RAG (text only)

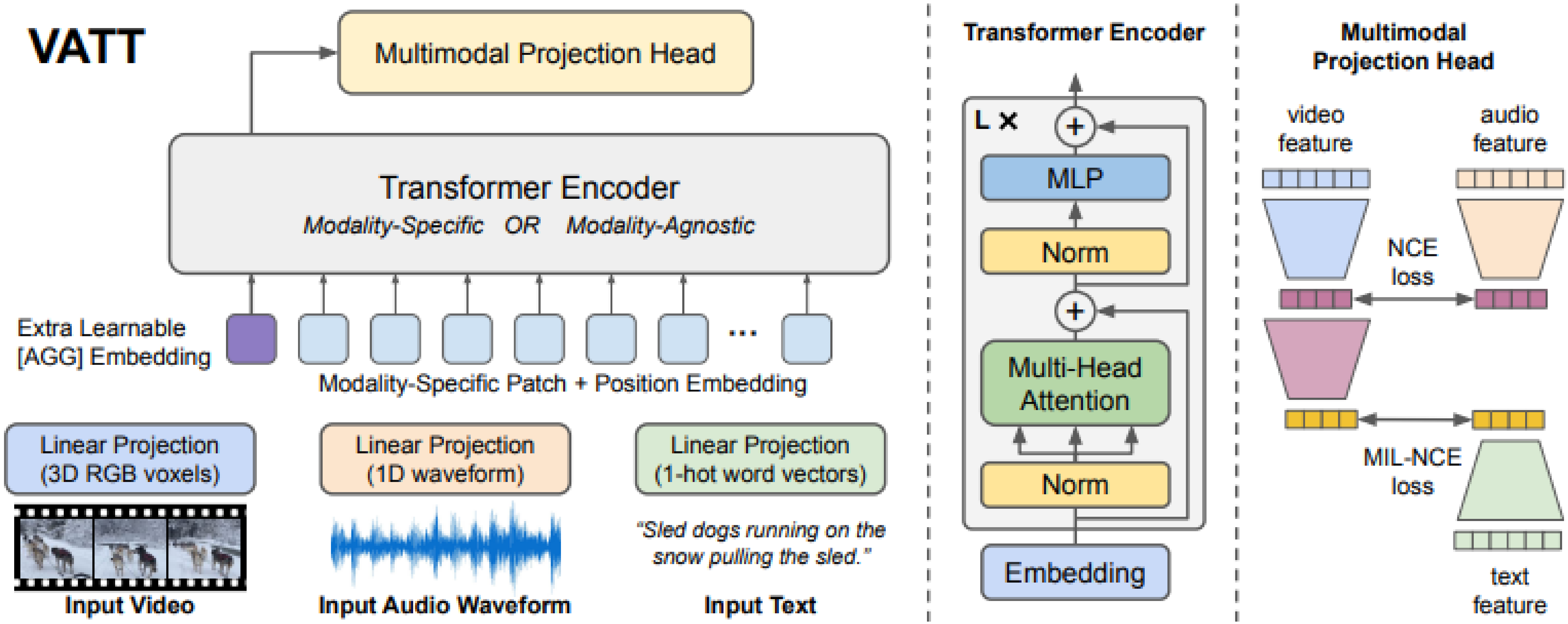


Humans communicate in multiple ways

Introducing Large Multimodal Models

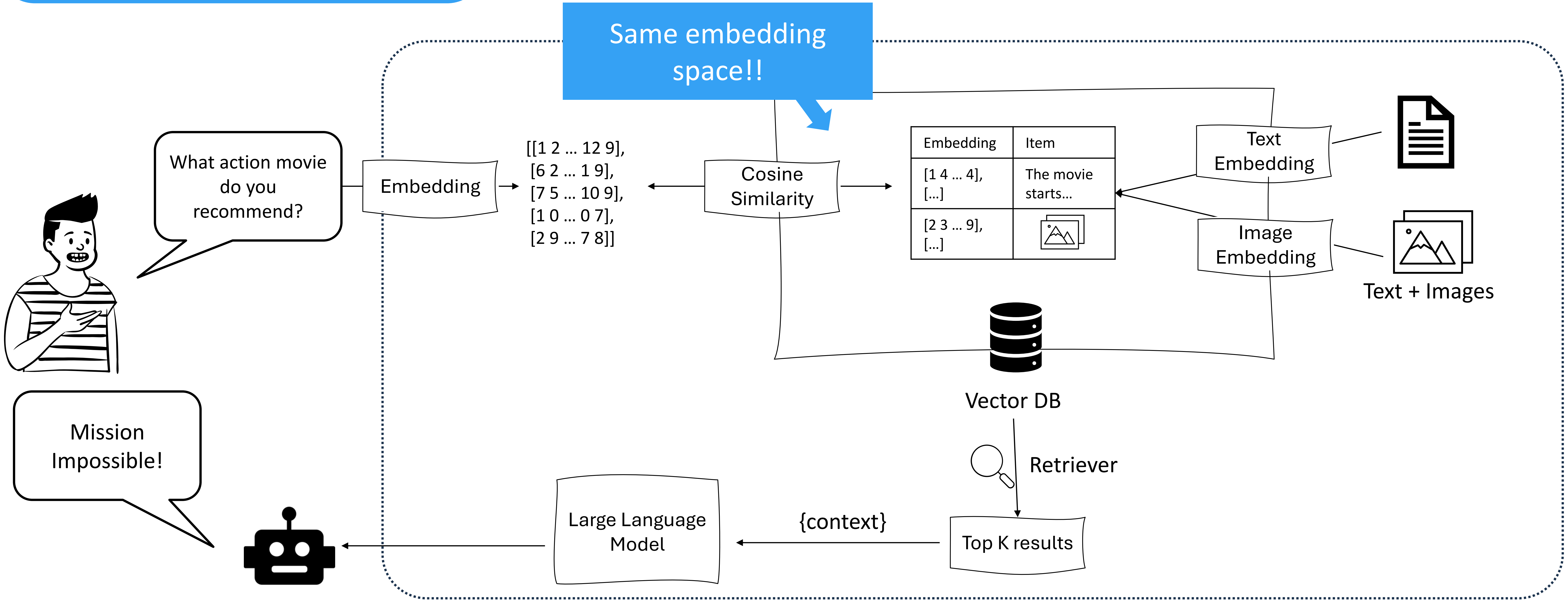


Example of Multimodality: VATT

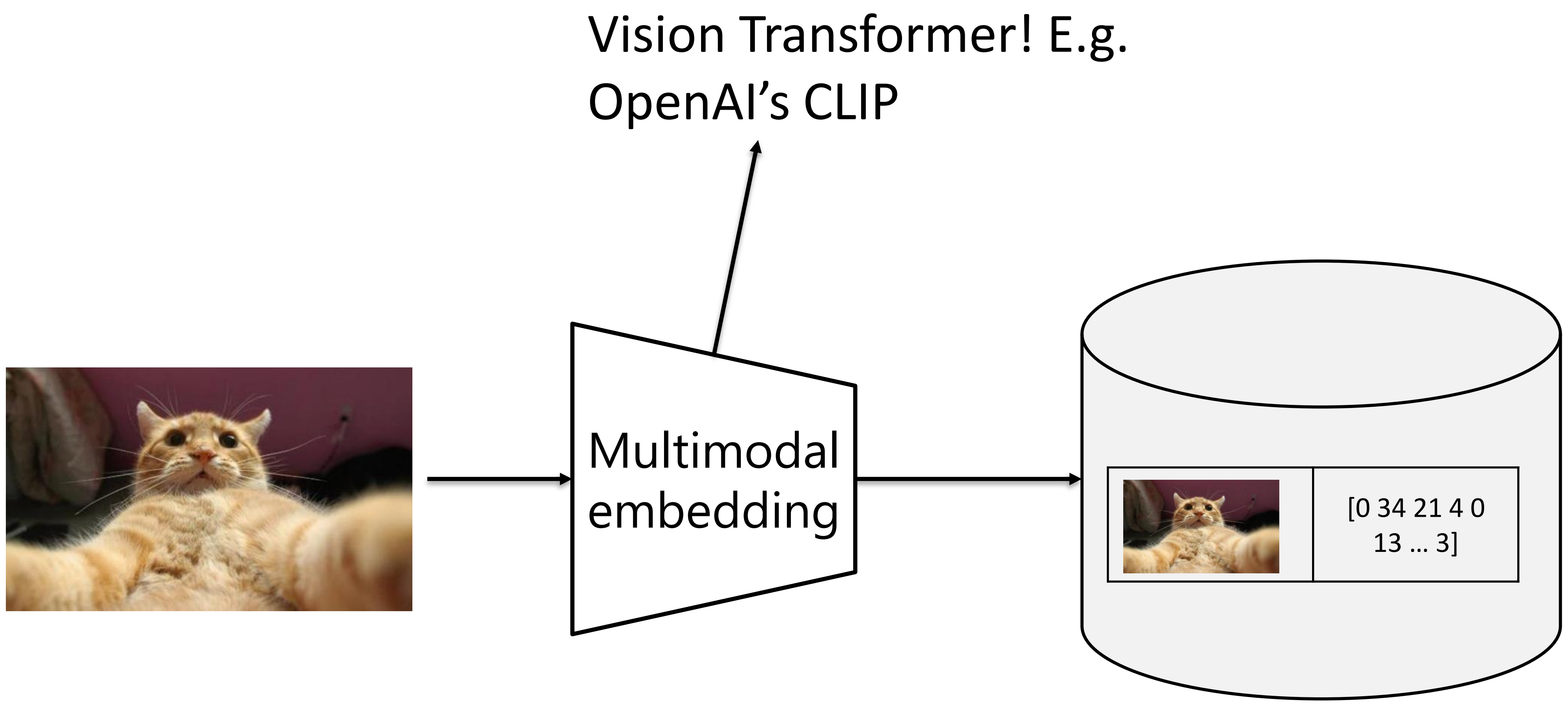


Source: *[2104.11178.pdf (arxiv.org)]*

Multimodal RAG (text + images)



Option 1: leverage an image embedding model



Attention is all you need – once more

AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Alexey Dosovitskiy[*,†], Lucas Beyer[*], Alexander Kolesnikov[*], Dirk Weissenborn[*], Xiaohua Zhai[*], Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby[*,†]

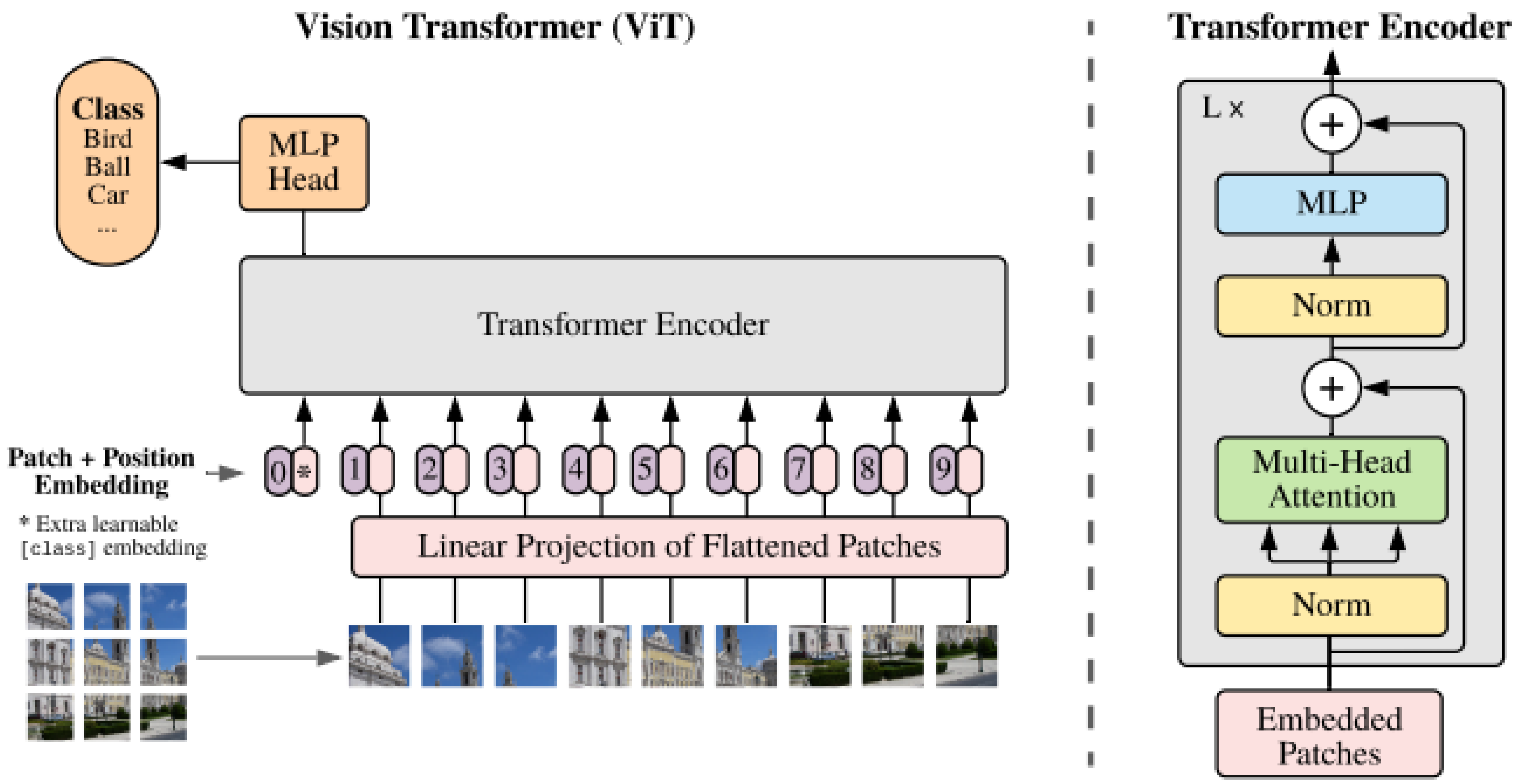
[*]equal technical contribution, [†]equal advising

Google Research, Brain Team

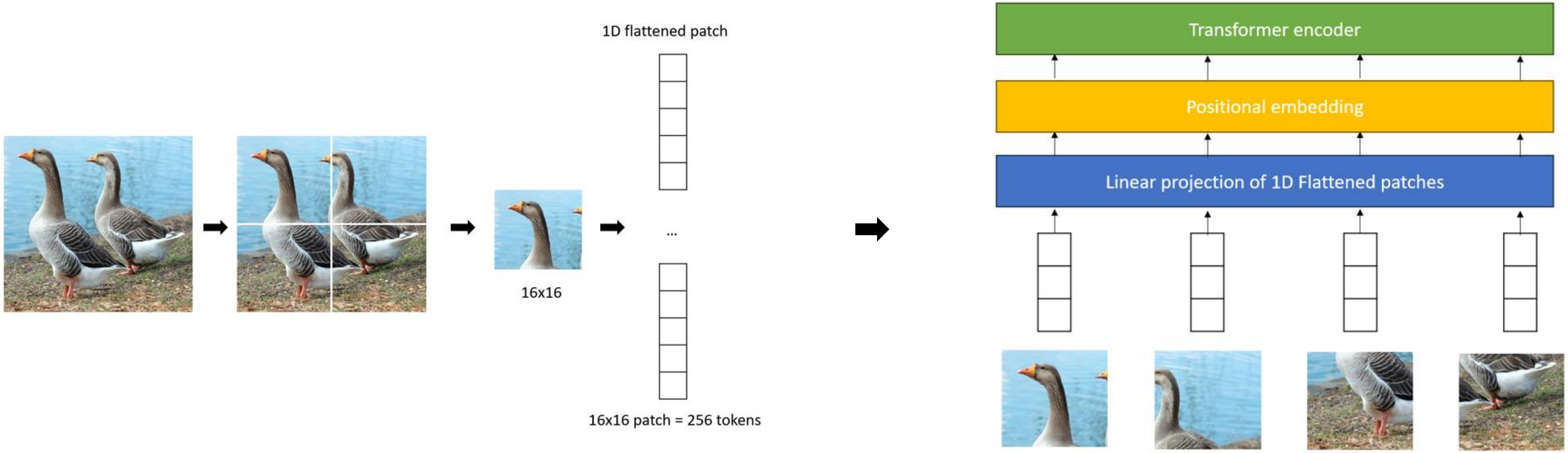
{adosovitskiy, neilhoulsby}@google.com

ABSTRACT

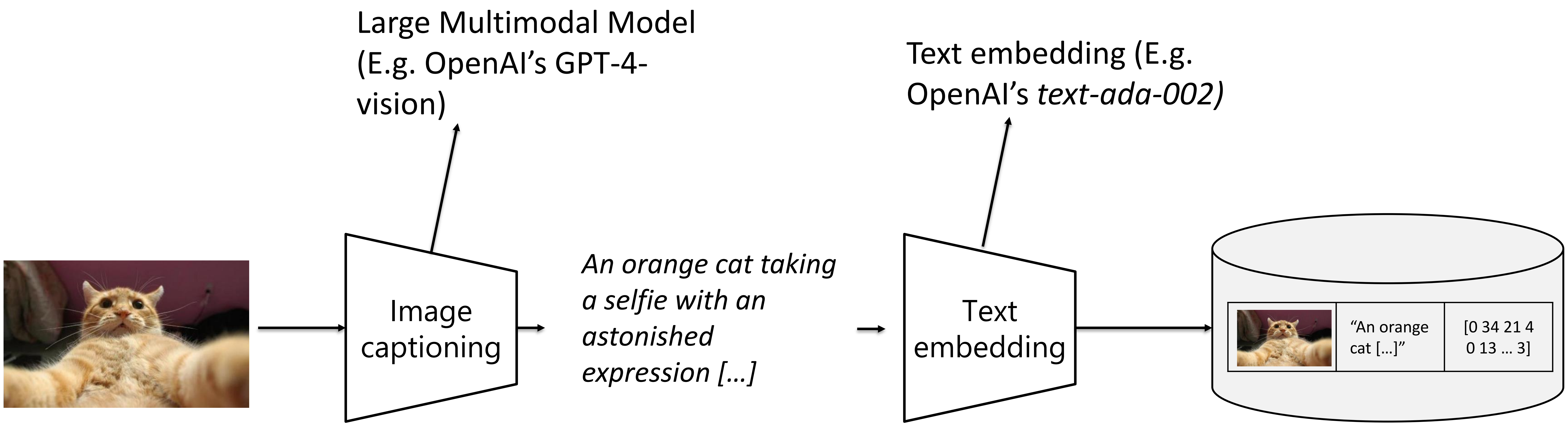
While the Transformer architecture has become the de-facto standard for natural language processing tasks, its applications to computer vision remain limited. In vision, attention is either applied in conjunction with convolutional networks, or used to replace certain components of convolutional networks while keeping their overall structure in place. We show that this reliance on CNNs is not necessary and a pure transformer applied directly to sequences of image patches can perform very well on image classification tasks. When pre-trained on large amounts of data and transferred to multiple mid-sized or small image recognition benchmarks (ImageNet, CIFAR-100, VTAB, etc.), Vision Transformer (ViT) attains excellent results compared to state-of-the-art convolutional networks while requiring substantially fewer computational resources to train.[1]



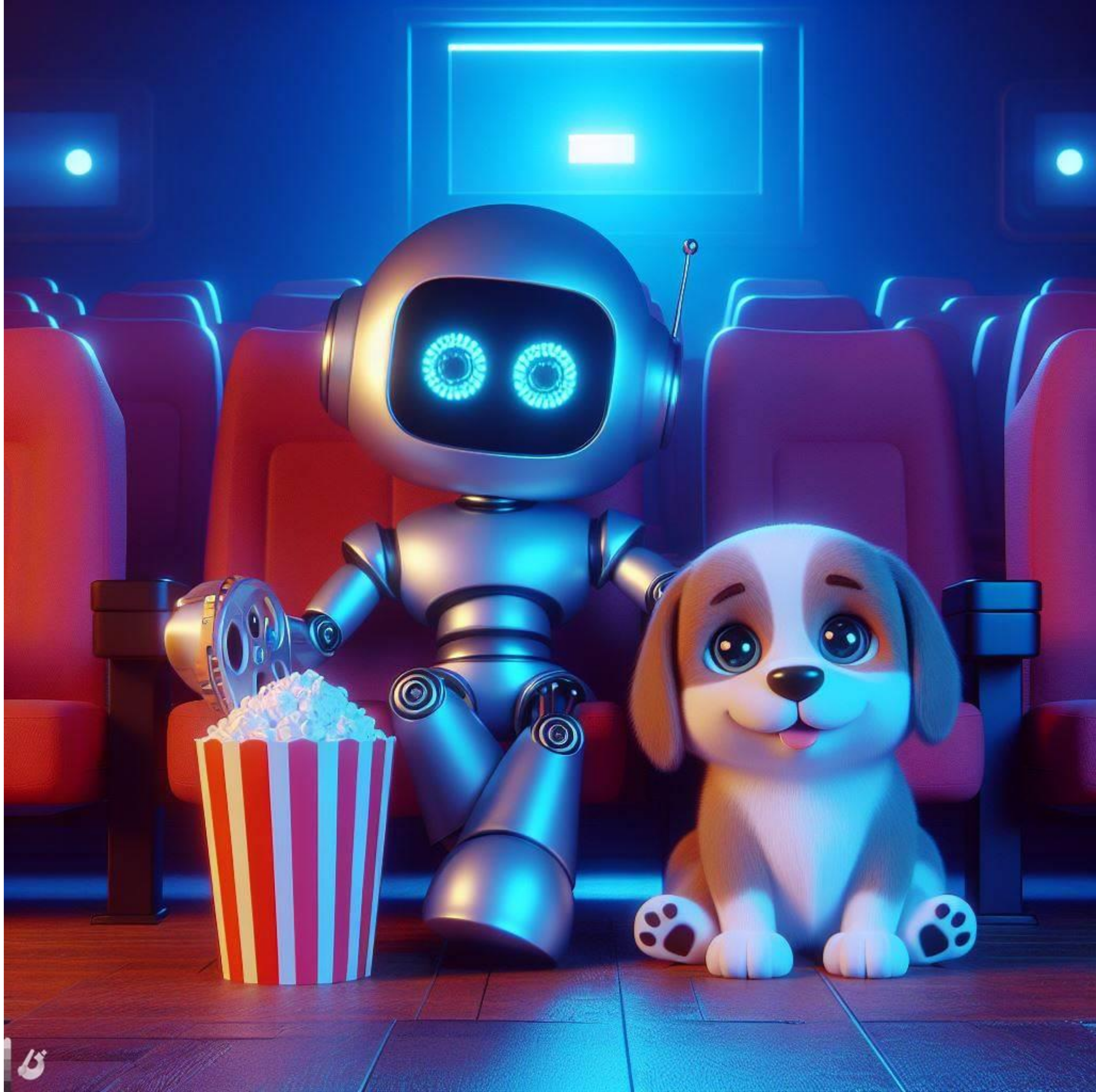
Breaking down images into patches



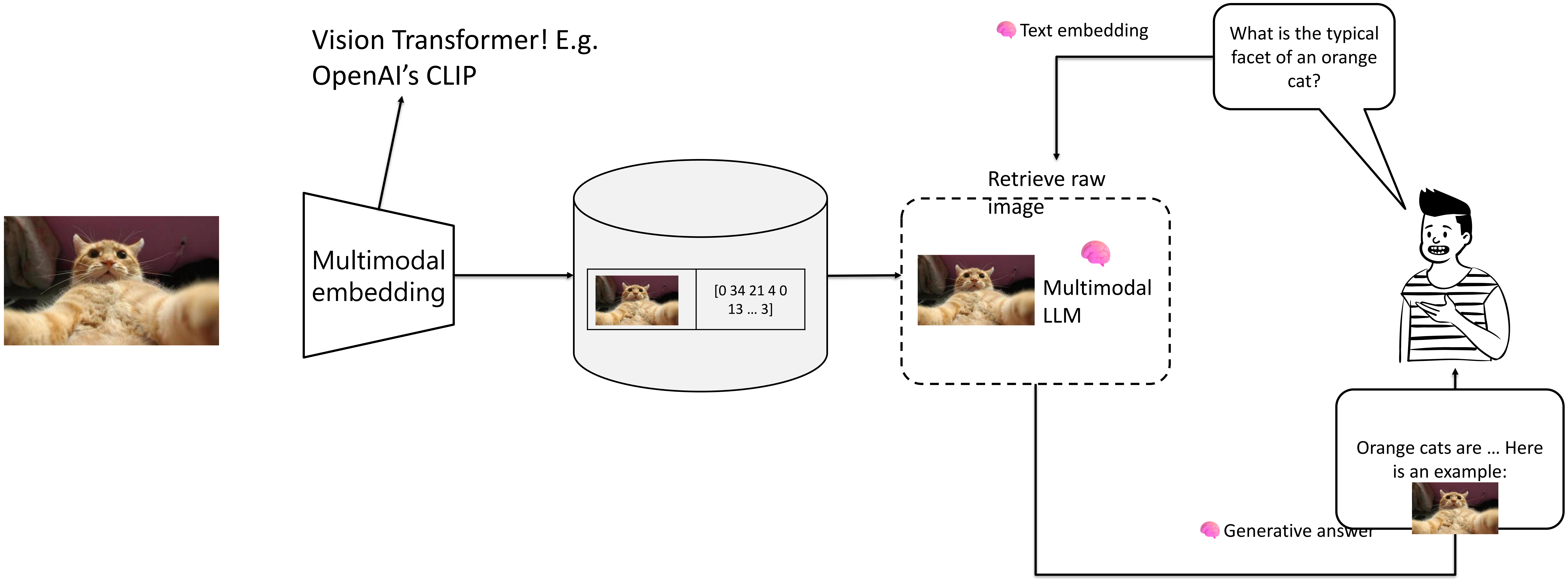
Option 2: Using an LMM to generate images embeddings



Demo Time!



Option 1: Image embeddings + text embeddings



Our Ingredients

MODELS

GPT-4
GPT4-TURBO-VISION

EMBEDDING

CLIP
TEXT-EMBEDDING-ADA-002

VECTORDB

QDRANT

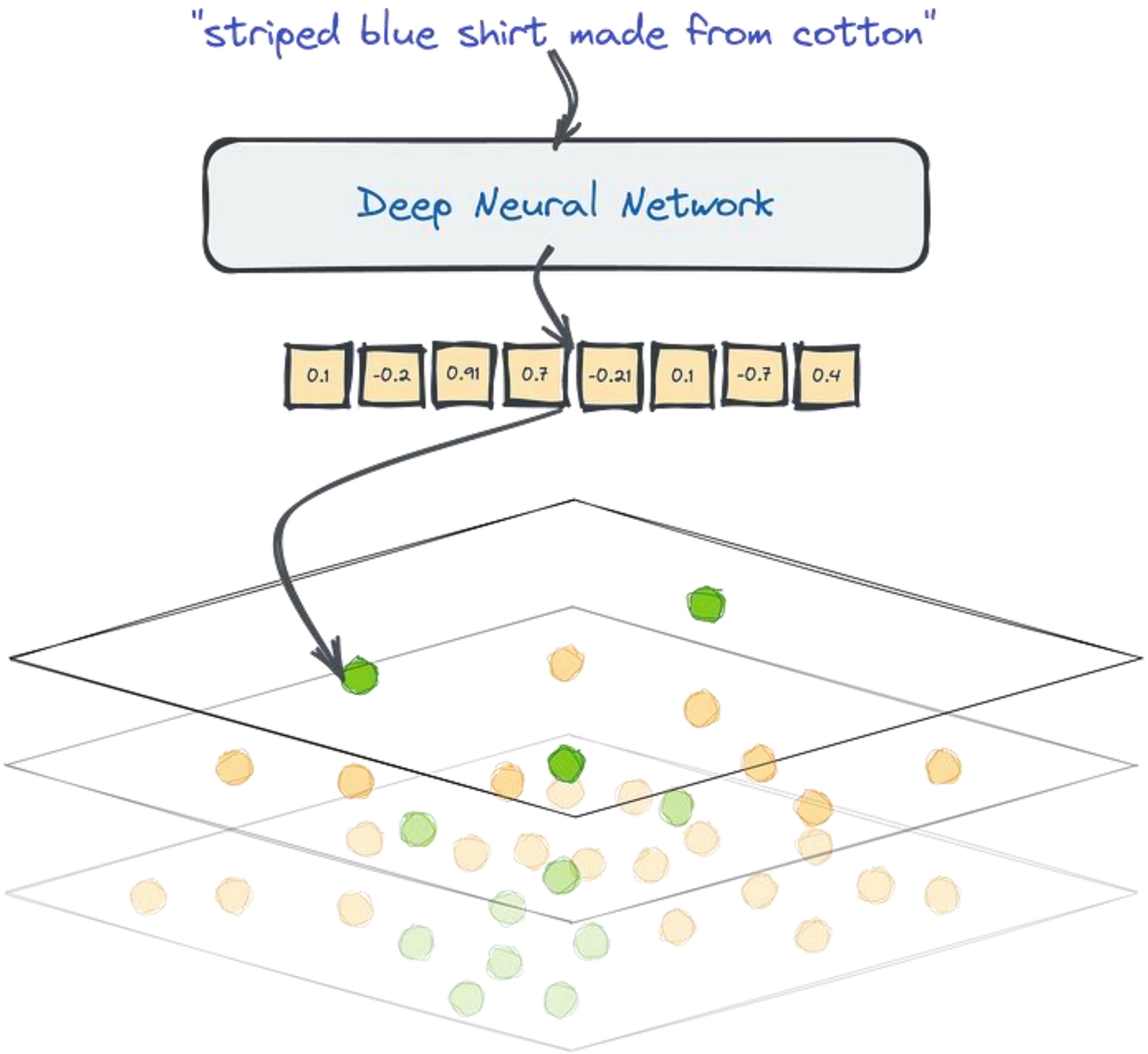
ORCHESTRATOR

LLAMA-INDEX

Qdrant

🚨 **Challenge→**find similar documents in a big set of objects

💡 **Solution→**using a graph-like structure to find the closest object, so that we compute the distance for some candidates only.



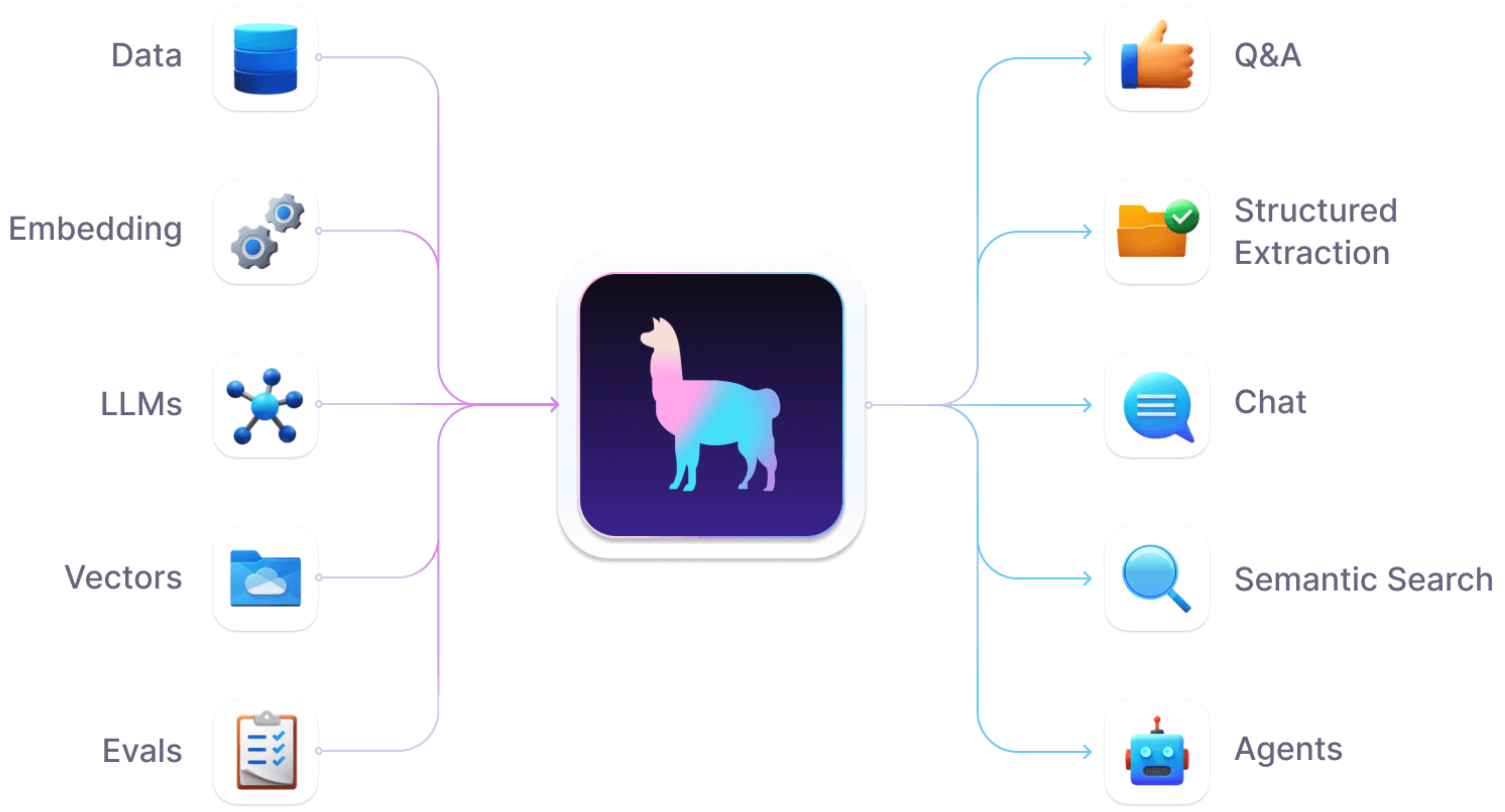
Source: *https://qdrant.tech/documentation/overview/vector-search/*

Llama-Index

Data Connectors to ingest data from various sources

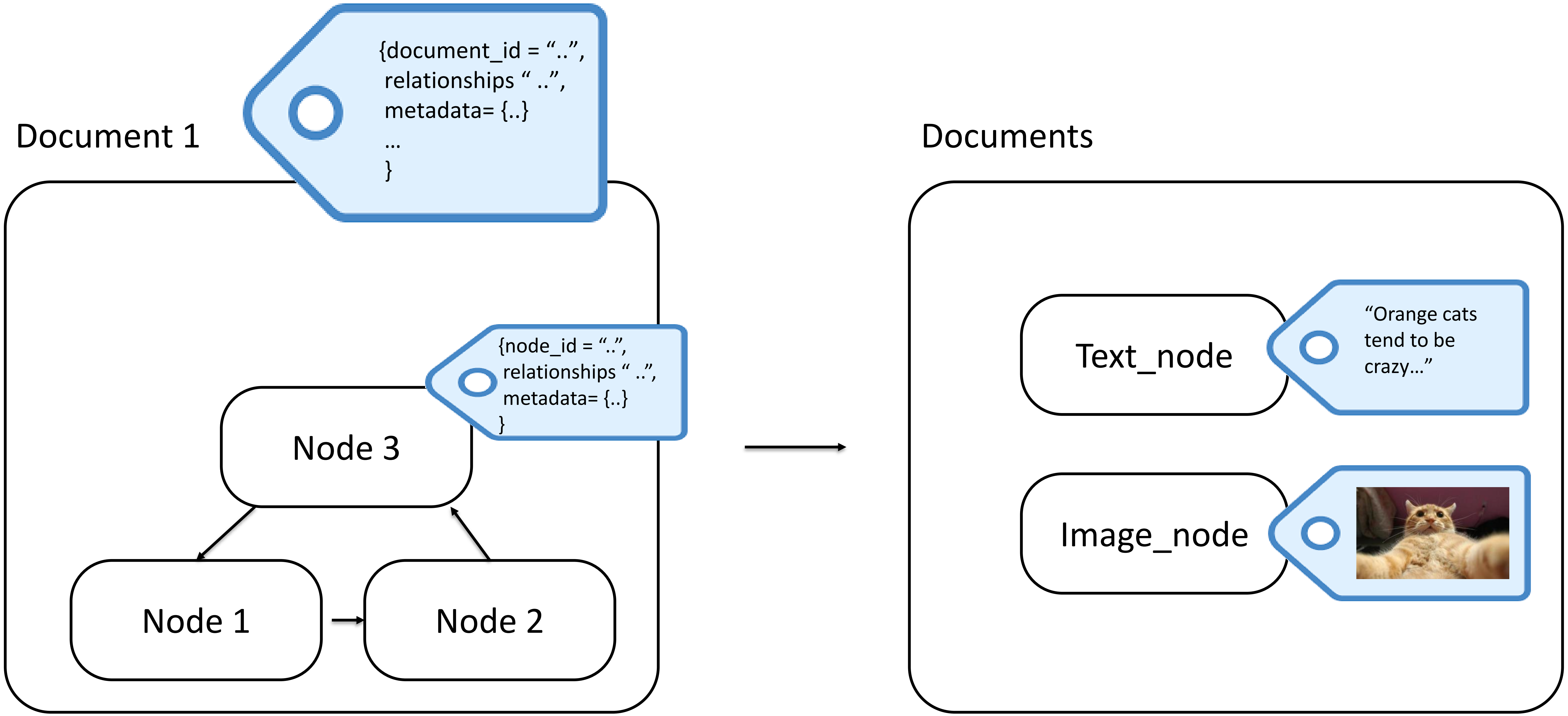
Data Indexes to structure your data in intermediate representations

Engines to interact with your data in natural language



Source: https://www.llamaindex.ai/

Nodes and Documents are first citizens in Llama-index



VALENTINA ALTO

APRIL 23, 2024

THANK YOU!

Let's Keep in touch!