

Applied Deep
Learning

5350 Graph Retrieval Augmented Generation (RAG)

portilla@gmail.com

Jesus.portilla@Colorado.edu

<https://github.com/iportilla/RAG>

Retrieval Augmented Generation (RAG)

USING LOCAL LLMS WITH LOCAL DATA

BASE MODEL

What was the name of that movie Chris emailed me about last year?



Sorry, I don't have access to any specific information about emails or ...

BASE MODEL + USER DATASET

What was the name of that movie Chris emailed me about last year?



Chris Thomson emailed you about "The Fall" on October 23rd last year. The Fall is an adventure fantasy film released in 2006, starring...



LLM Shortcomings

write a `get_embedding` function in python using the latest azure openai client


python

 Copy

```
def get_embedding(texts, model="text-embedding-ada-002"):
    """
    Retrieve embeddings for a list of texts using the specified model.

    Parameters:
    - texts (list of str): The texts to embed.
    - model (str): The model to use for embedding. Default is "text-embedding-ada-002".

    Returns:
    - embeddings (list of list of floats): The embeddings for each text.
    """
    response = openai.Embedding.create(
        input=texts,
        model=model
    )
    embeddings = [embedding['embedding'] for embedding in response['data']]
    return embeddings
```



Incorporate Domain Knowledge with LLM



Prompt engineering

In-context learning



Fine tuning

Learn new skills
(permanently)



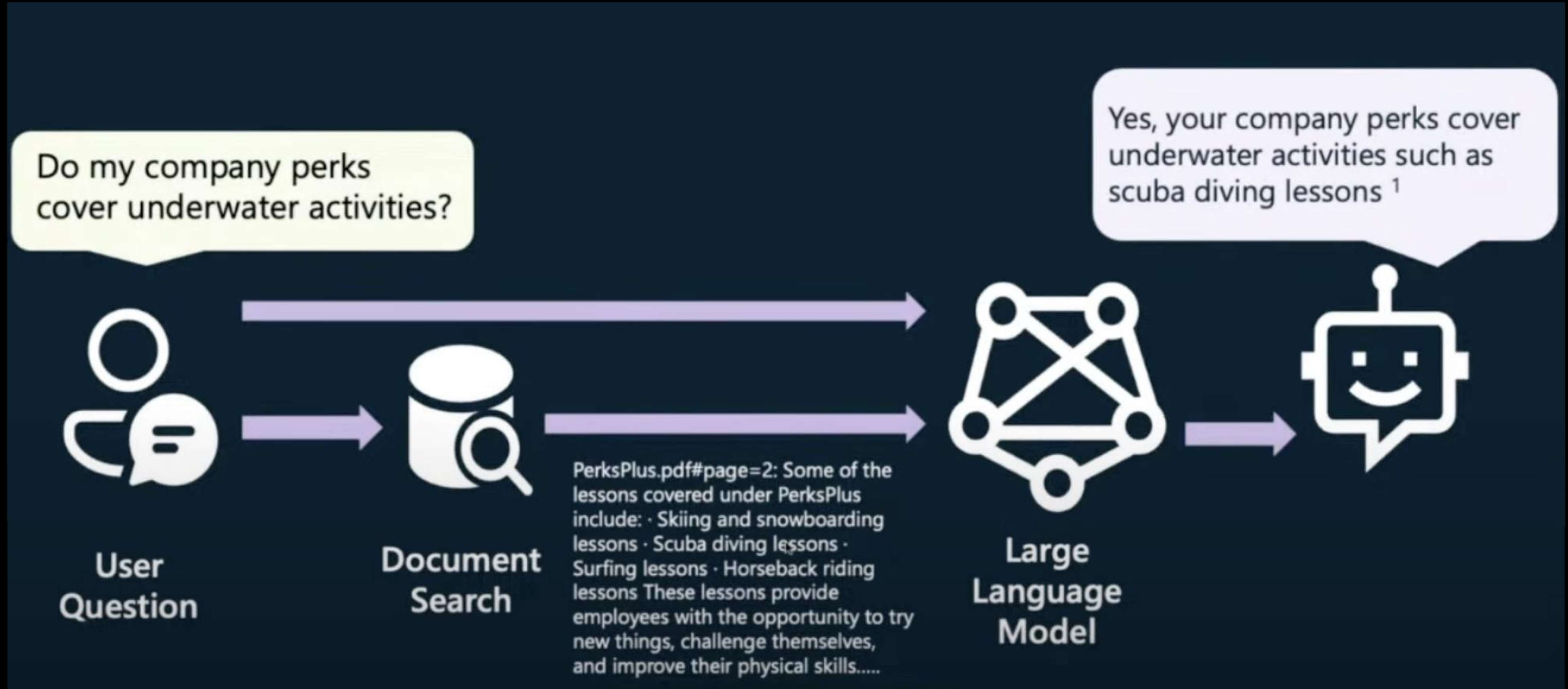
Retrieval augmentation

Learn new facts
(temporarily)

The Benefits of RAG

- Up-to-date public knowledge (AZ OpenAI documentation)
- Access to internal knowledge (Company HR docs)

RAG – Retrieval Augmented Generation



Robust retrieval for RAG

- Responses only as good as retrieved data
- Keyword search recall challenges
- Vector-based retrieval finds docs by **Semantic** similarity

Example

Question:

"Looking for lessons on underwater activities"



Won't match:

"Scuba classes"

"Snorkeling group sessions"

Vector embeddings

- An embedding encodes an input as a list of FP numbers
- “dog” -> [0.014, -0.05, ...]
- Different models output different embeddings (different lengths)

<https://aka.ms/aitour/vectors>

<https://pamelafox.github.io/vectors-comparison/>

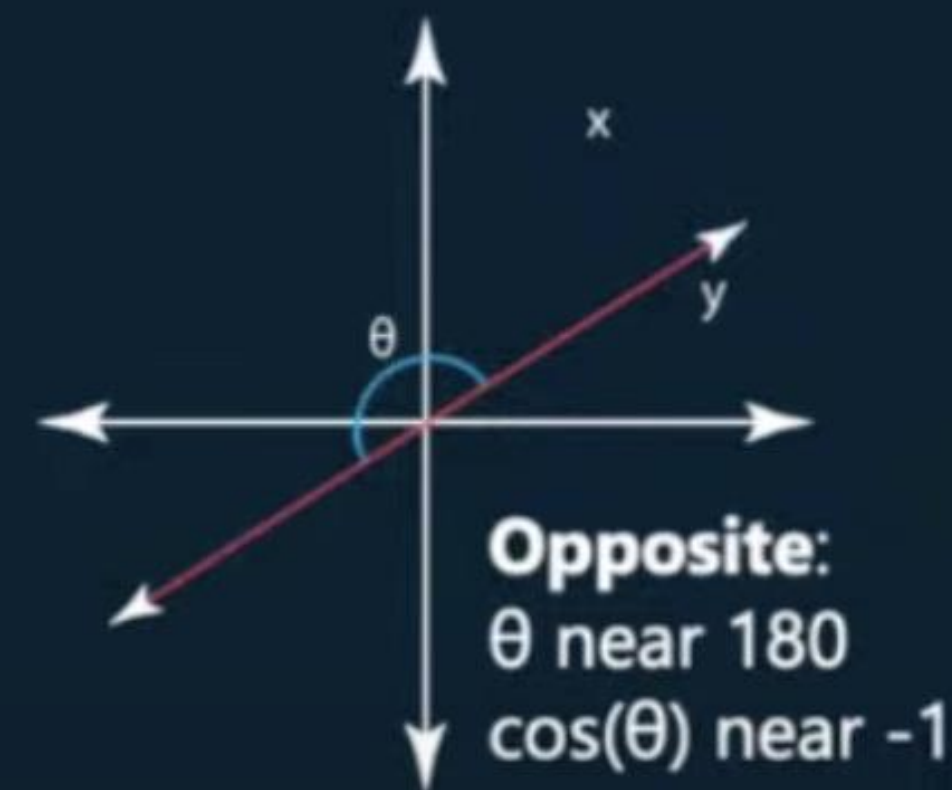
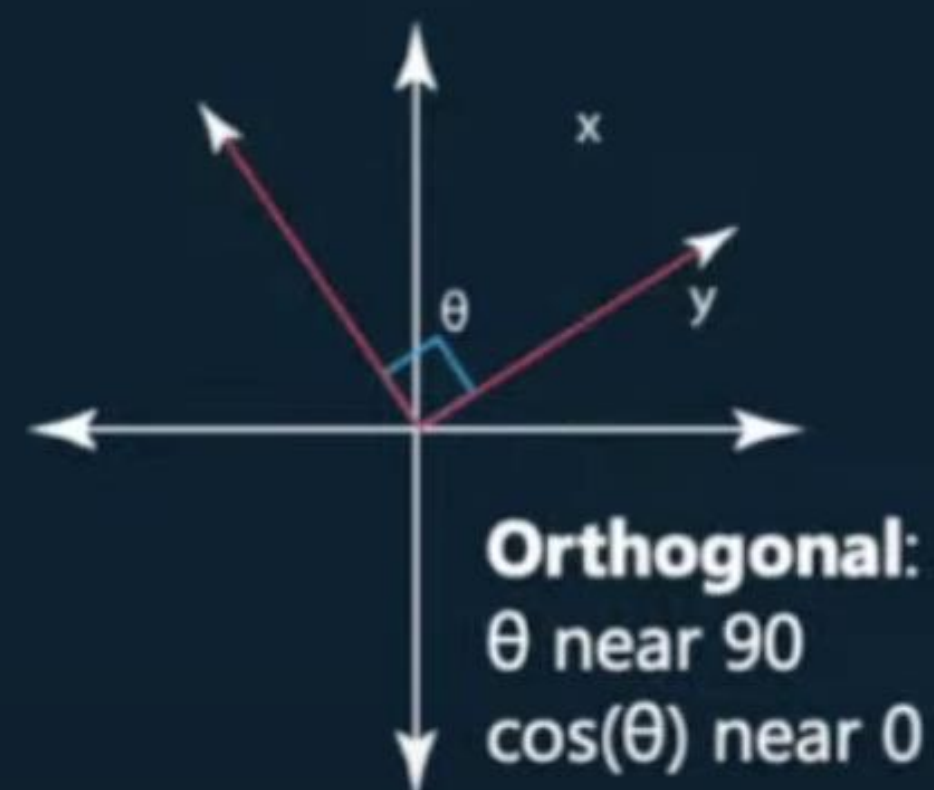
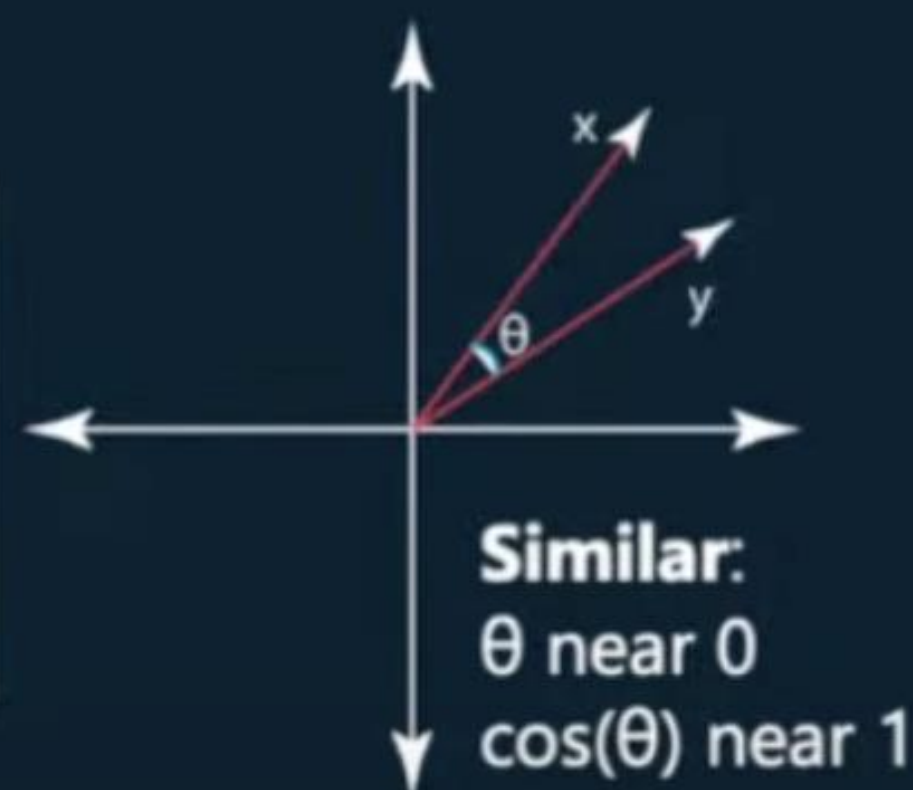
<https://pamelafox.github.io/vectors-comparison/movies.html>

https://github.com/Azure-Samples/rag-with-azure-ai-search-notebooks/blob/main/vector_embeddings.ipynb

Vector similarity

Embeddings are used to calculate similarity between inputs:
The most common distance measurement is cosine similarity

```
def cosine_sim(a, b):  
    return dot(a, b) /  
        (mag(a) * mag(b))
```

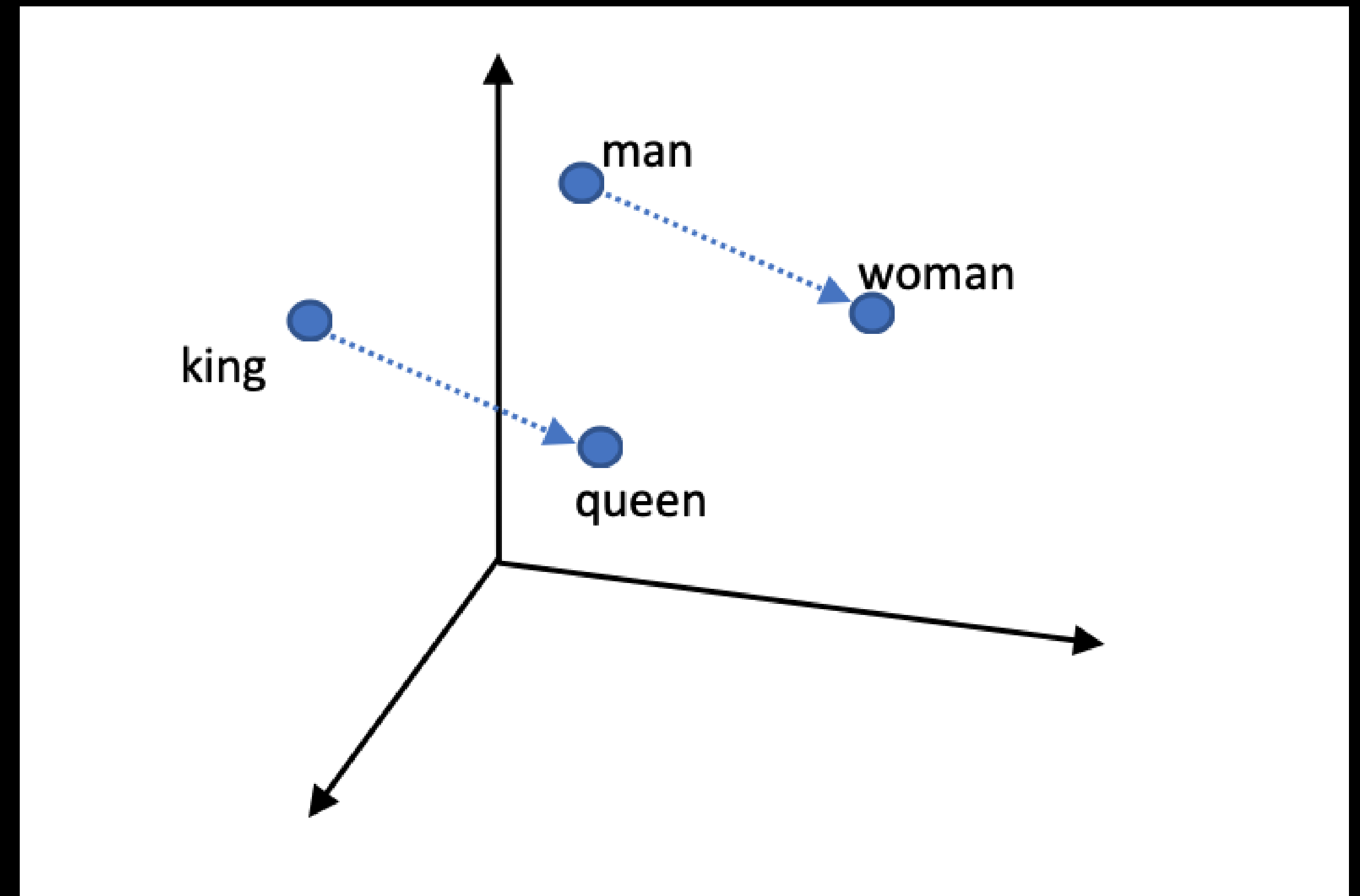
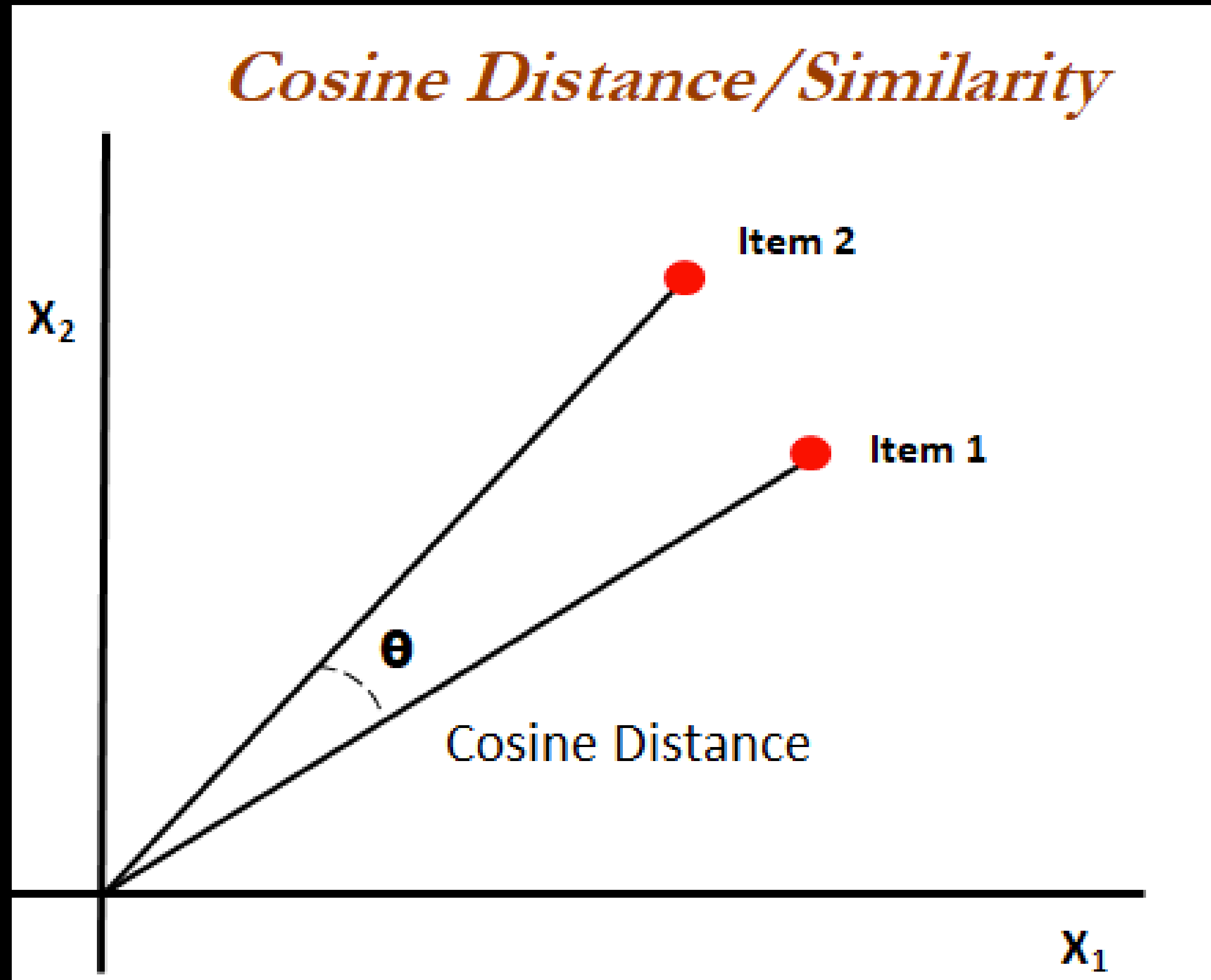


*For ada-002, $\cos(\theta)$ values range from 0.7-1

<https://aka.ms/aitour/vectors>

https://github.com/Azure-Samples/rag-with-azure-ai-search-notebooks/blob/main/vector_embeddings.ipynb

Vector embeddings



Vector Comparison

What is a vector?

Expore words from a dataset of 1000 words across two embedding models.

Target word:

Embedding model:

Both (Comparison) ▾

Find word

| Model: word2vec | Model: openai | | | | | | | | | | | | | | | | | | | | | | | | |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------|--------------------|-----------------------|--------------------|-----------------------|---------------------|--|--|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------|--------------------|-----------------------|--------------------|----------------------|--------------------|------------------------|--------------------|------------------------|--------------------|------------------------|--------------------|----------------------|--------------------|--|--|
| <div>Vector: 300 dimensions</div> <div><div>0.044865, -0.010391, -0.017868, 0.027773, 0.055935, 0.01209, -0.017383, 0.097498, 0.034765, -0.020102, 0.09206, -0.029716, 0.08701, 0.01379, -0.057878, 0.022918, 0.002671, -0.002792, 0.052439, -0.100994, 0.057101, -0.055935, -0.014178, -0.08468, -0.098664, 0.01981, -0.036125, 0.057489, 0.022724, -0.041369, -0.078076, -0.081572, -0.10954, 0.012187, 0.080019, 0.069142, 0.036319, -0.040204, 0.090895, -0.016217, 0.010779, -0.000422, 0.010779, 0.135954, -0.052439</div><div>Most similar:</div><table><tr><td>read</td><td>0.3893648604097623</td></tr><tr><td>paper</td><td>0.3634623893904801</td></tr><tr><td>write</td><td>0.35940013889130784</td></tr><tr><td></td><td></td></tr></table></div> | read | 0.3893648604097623 | paper | 0.3634623893904801 | write | 0.35940013889130784 | | | <div>Vector: 1536 dimensions</div> <div><div>-0.006843345705419779, -0.019184302538633347, -0.004917495418339968, -0.022664999589323997,</div><div>Most similar:</div><table><tr><td>paper</td><td>0.8874017308879492</td></tr><tr><td>movie</td><td>0.8805337935966647</td></tr><tr><td>film</td><td>0.8711653176455576</td></tr><tr><td>letter</td><td>0.8632871648170634</td></tr><tr><td>record</td><td>0.8630170946356468</td></tr><tr><td>course</td><td>0.8629488396382509</td></tr><tr><td>bank</td><td>0.8628000814561154</td></tr><tr><td></td><td></td></tr></table></div> | paper | 0.8874017308879492 | movie | 0.8805337935966647 | film | 0.8711653176455576 | letter | 0.8632871648170634 | record | 0.8630170946356468 | course | 0.8629488396382509 | bank | 0.8628000814561154 | | |
| read | 0.3893648604097623 | | | | | | | | | | | | | | | | | | | | | | | | |
| paper | 0.3634623893904801 | | | | | | | | | | | | | | | | | | | | | | | | |
| write | 0.35940013889130784 | | | | | | | | | | | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | | | | | | | | | | | |
| paper | 0.8874017308879492 | | | | | | | | | | | | | | | | | | | | | | | | |
| movie | 0.8805337935966647 | | | | | | | | | | | | | | | | | | | | | | | | |
| film | 0.8711653176455576 | | | | | | | | | | | | | | | | | | | | | | | | |
| letter | 0.8632871648170634 | | | | | | | | | | | | | | | | | | | | | | | | |
| record | 0.8630170946356468 | | | | | | | | | | | | | | | | | | | | | | | | |
| course | 0.8629488396382509 | | | | | | | | | | | | | | | | | | | | | | | | |
| bank | 0.8628000814561154 | | | | | | | | | | | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | | | | | | | | | | | |

Movie title embeddings in OpenAI

Movie title embeddings in OpenAI

Expore embeddings for Disney movie titles from OpenAI ada-002 model.

Select a movie title:

See embedding

Movie title: The Jungle Book

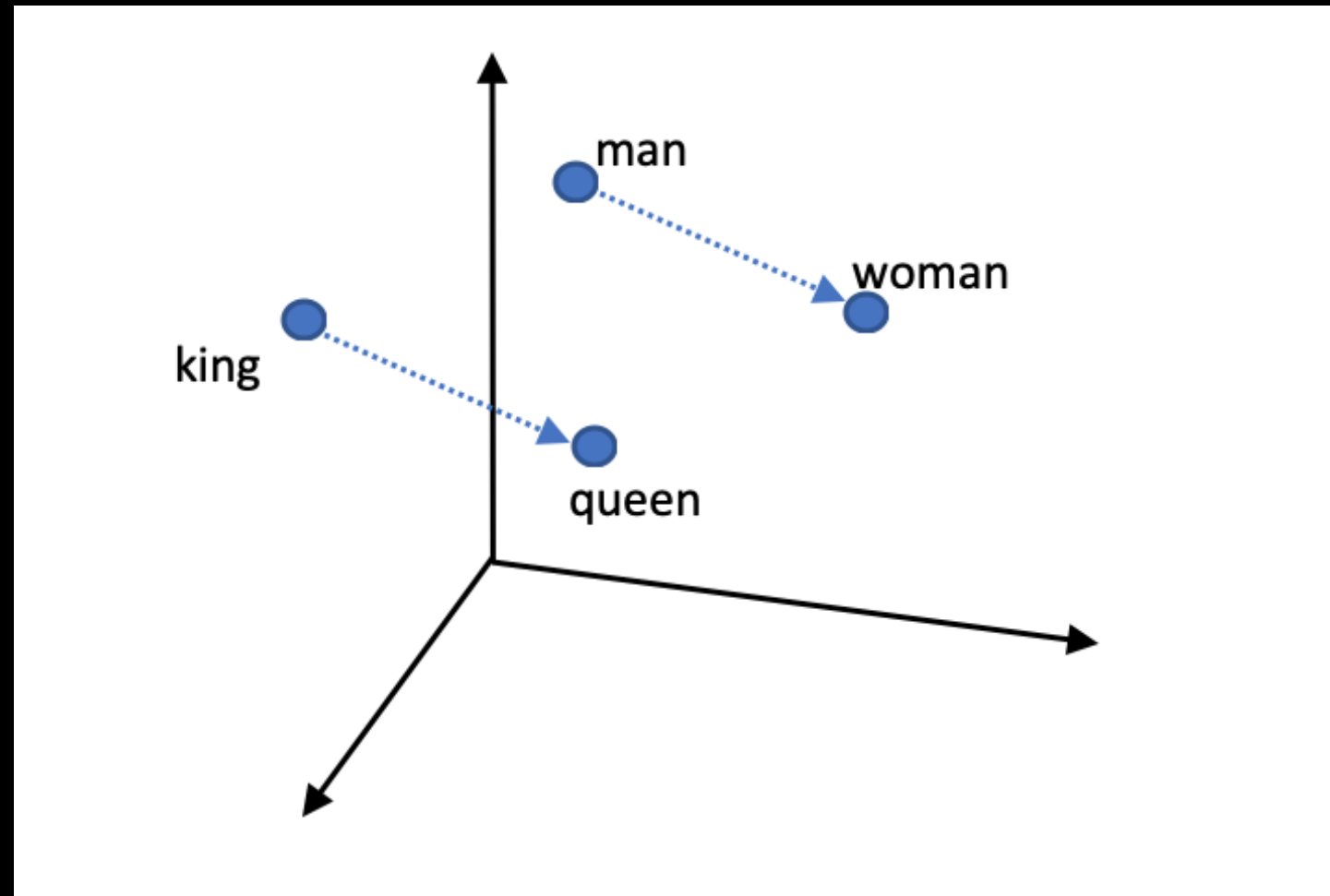
Vector: 1536 dimensions

-0.009433940052986145, -0.0026398864574730396, 0.002852880861610174, -0.0006918430444784462, -0.01920369639992714, 0.017636556178331375, -0.013955017551779747, -0.024390187114477158,

Most similar:

| | |
|---------------------------------------|--------------------|
| The Jungle Book 2 | 0.9486278980316131 |
| Jungle 2 Jungle | 0.9236481731450379 |
| The Lion King | 0.9001141316128429 |
| George Of The Jungle | 0.8967382582947568 |
| Tarzan | 0.8928694263214043 |
| The Fox and the Hound | 0.8667384685848213 |
| The Tigger Movie | 0.8659348715821917 |

Vector embeddings Lab



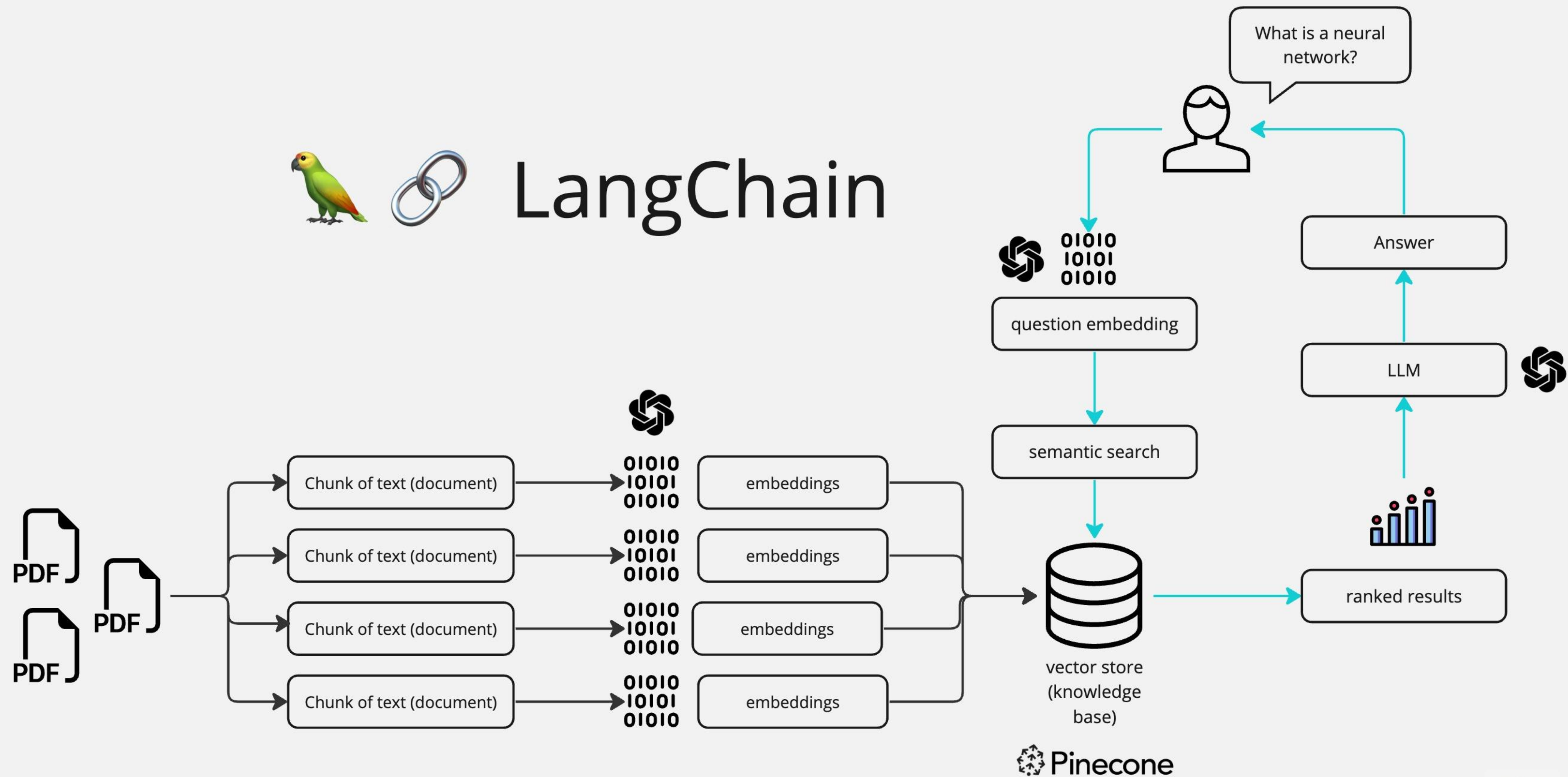
<https://aka.ms/aitour/vectors>

- Azure OpenAI
- RAG
- Exercise

RAG

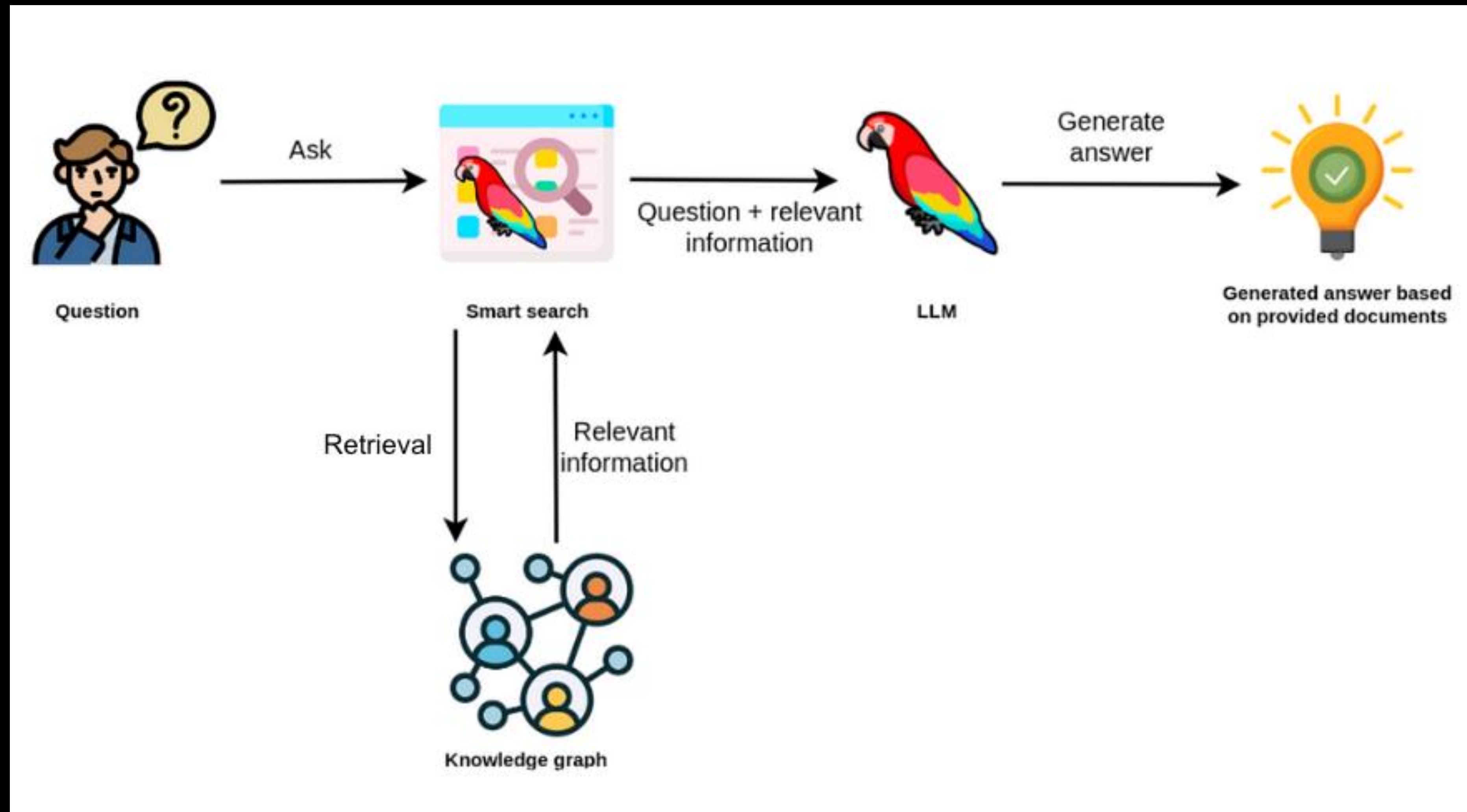


LangChain

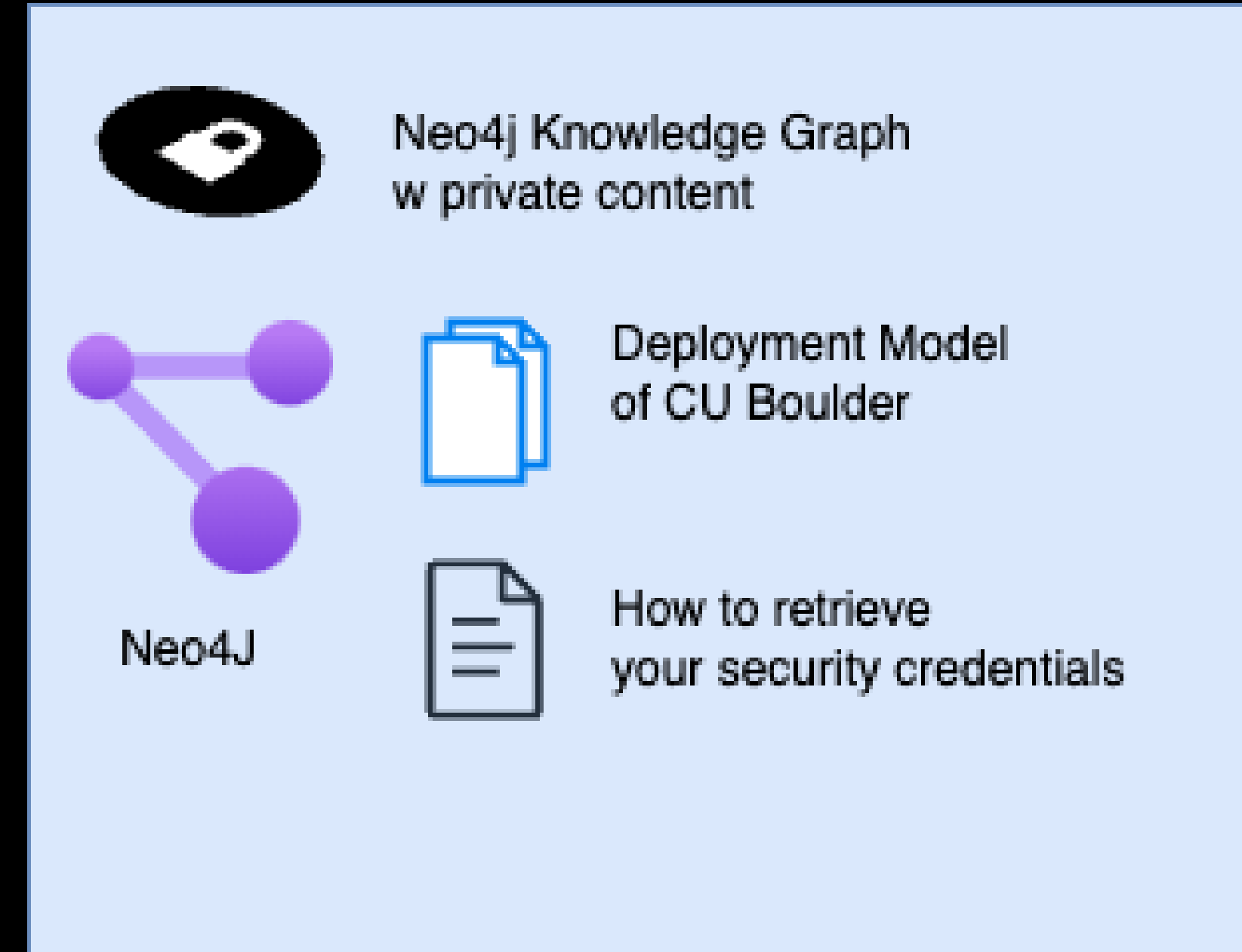
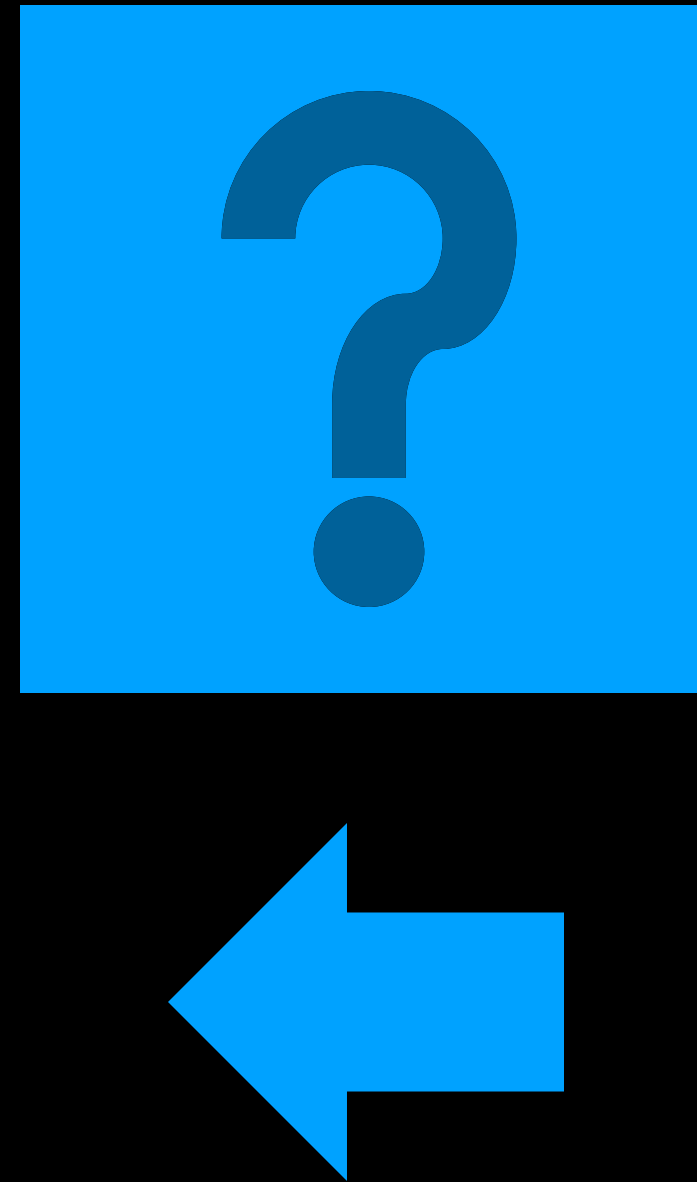
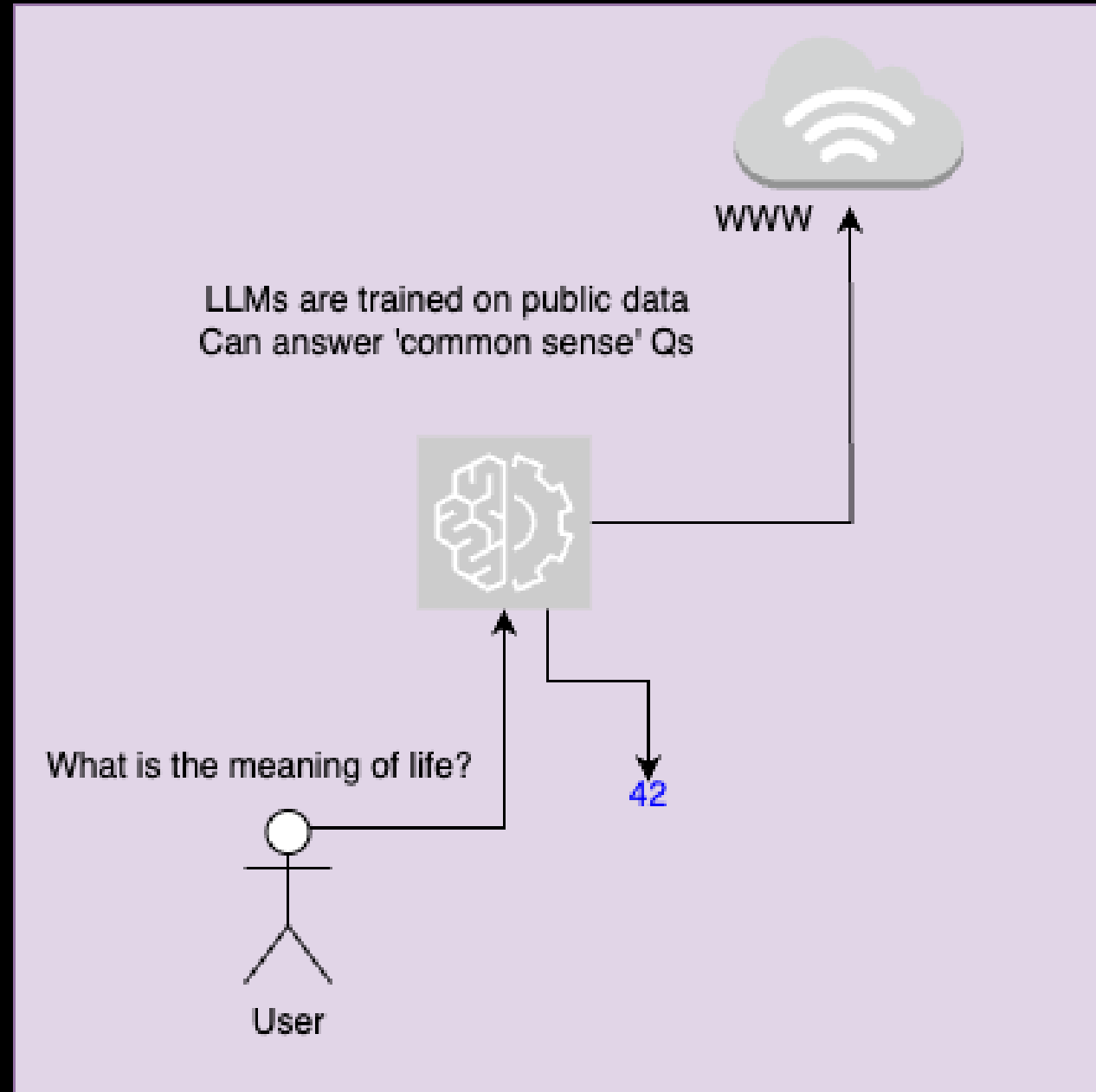


miro

Graph RAG



Graph RAG



Graph-RAG Lab

powered by Neo4j

Chatbot

Which manager owns FAANG stocks?

- VANGUARD GROUP INC: 6,104,763,905 shares, \$539,729,670,719,000
- BlackRock Inc.: 4,953,610,092 shares, \$435,415,791,674,000
- STATE STREET CORP: 3,740,913,030 shares, \$353,538,102,181,000
- Berkshire Hathaway Inc: 2,737,381,876 shares, \$269,391,573,936,000
- FMR LLC: 1,928,066,413 shares, \$161,924,426,418,000
- NORTHERN TRUST CORP: 1,218,471,570 shares, \$96,774,371,692,000
- MORGAN STANLEY: 1,007,307,578 shares, \$89,984,740,936,000
- Arete Wealth Advisors, LLC: 998,925 shares, \$86,297,215,782,000
- BANK OF AMERICA CORP /DE/: 839,135,819 shares, \$71,704,246,692,000
- GEODE CAPITAL MANAGEMENT, LLC: 1,356,471,688 shares, \$66,889,948,693,000

Examples

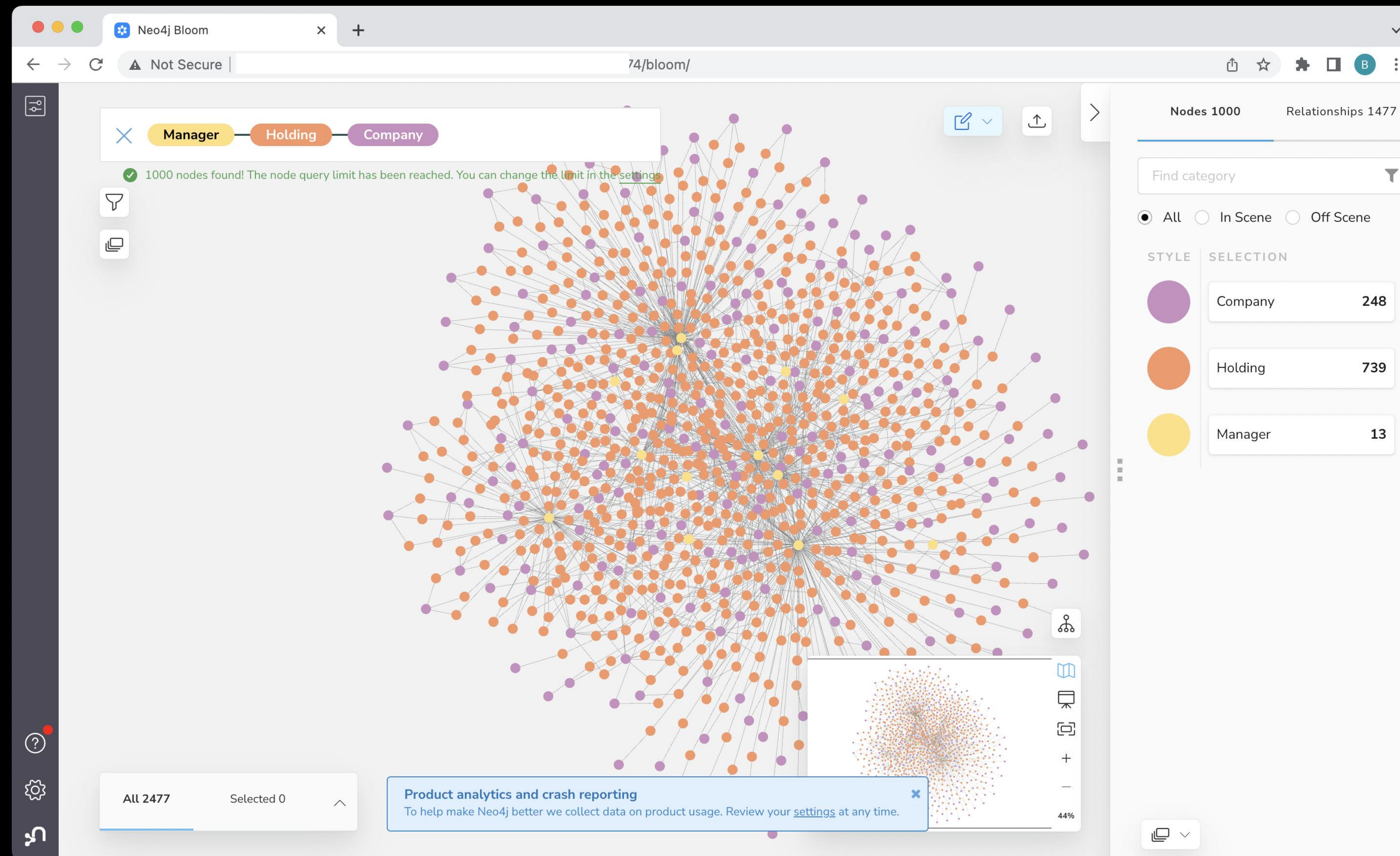
What are the top 10 investments for Vanguard? Which manager owns FAANG stocks? What are other top investments for fund managers investing in Exxon?

What are Rempart's top investments by value for 2023? Who are the common investors between Tesla and Costco?

Use via ODT Built with Gradio

<https://github.com/iportilla/genai-stack>

Graph-RAG Lab



<https://github.com/neo4j-partners/hands-on-lab-neo4j-and-azure>