# ABB - Session 4

## Software 3.0, RAG, & Text Embeddings

**Shaw Talebi**

ABB #3 - Winter 2025

# Today's Session

ABB #3 - Winter 2025

# Announcements

1) Mid-course Survey

ABB #3 - Winter 2025

# Homework 3
## Shoutouts 🎉

**AI Executive Coach**

Andry Haryanto

**Travel Itinerary Planner**

Kader Mula

**LLM-powered Anomaly Detection**

Shaji Ravindra Nathan

**Automated AI Newsletter**

Ravindra Bharathi

# 2 Levels of LLM Development

How to get LLMs to do what you want…

## Level 1

*Adapting models via prompts and tools*

Prompt Engineering 💬

RAG 🔍

Tool-use 🛠️

## Level 2

*Adapting models via additional training*

Fine-tuning 🪛

Post-training 👨‍💻
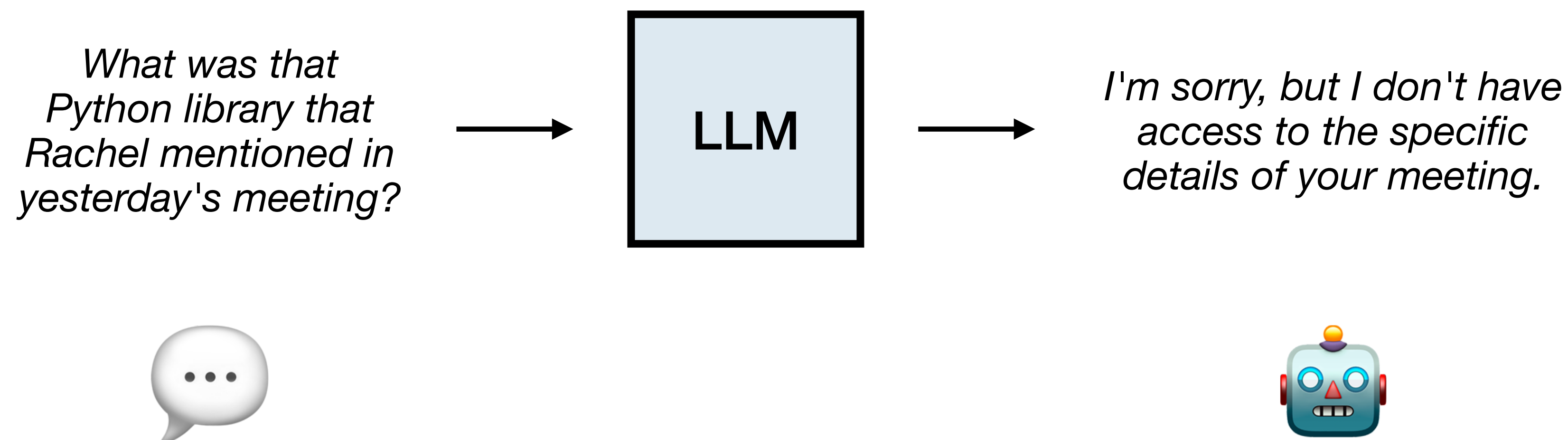
ABB #3 - Winter 2025

# RAG

ABB #3 - Winter 2025

# What is RAG?

Improving an LLM's responses by automatically providing it relevant context

## *RAG = Retrieval Augmented Generation*

*What was that Python library that Rachel mentioned in yesterday's meeting?* → LLM → *I'm sorry, but I don't have access to the specific details of your meeting.*

# What is RAG?

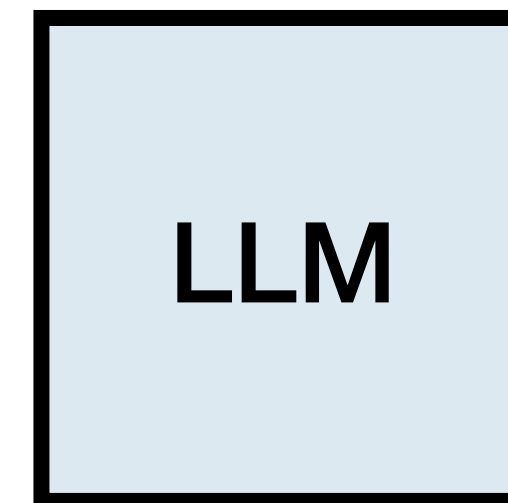Improving an LLM's responses by automatically providing it relevant context

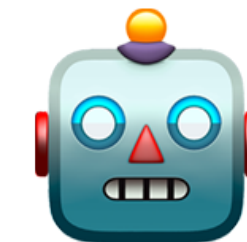*What was that Python library that Rachel mentioned in yesterday's meeting?*
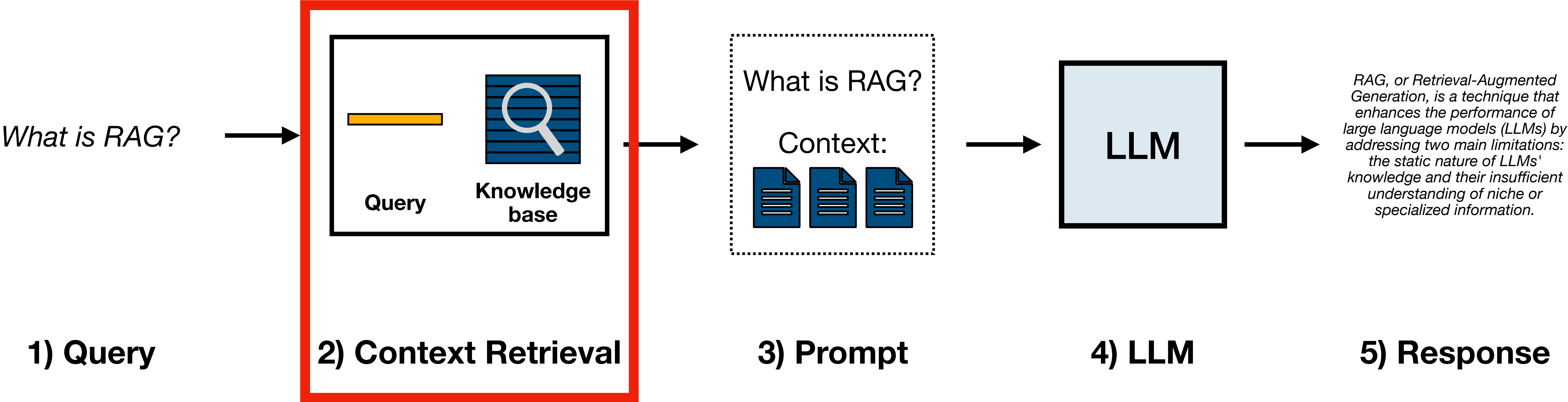
**Attach meeting transcript**

LLM

*PyMuPDF*

# RAG Implementation
## 5-step process

What is RAG? →



**1) Query**    **2) Context Retrieval**    **3) Prompt**    **4) LLM**    **5) Response**

Query    Knowledge base

What is RAG?

Context:

LLM

*RAG, or Retrieval-Augmented Generation, is a technique that enhances the performance of large language models (LLMs) by addressing two main limitations: the static nature of LLMs' knowledge and their insufficient understanding of niche or specialized information.*

# 2 Ways Retrieval Approaches

## Way 1

*Keyword-based Search*
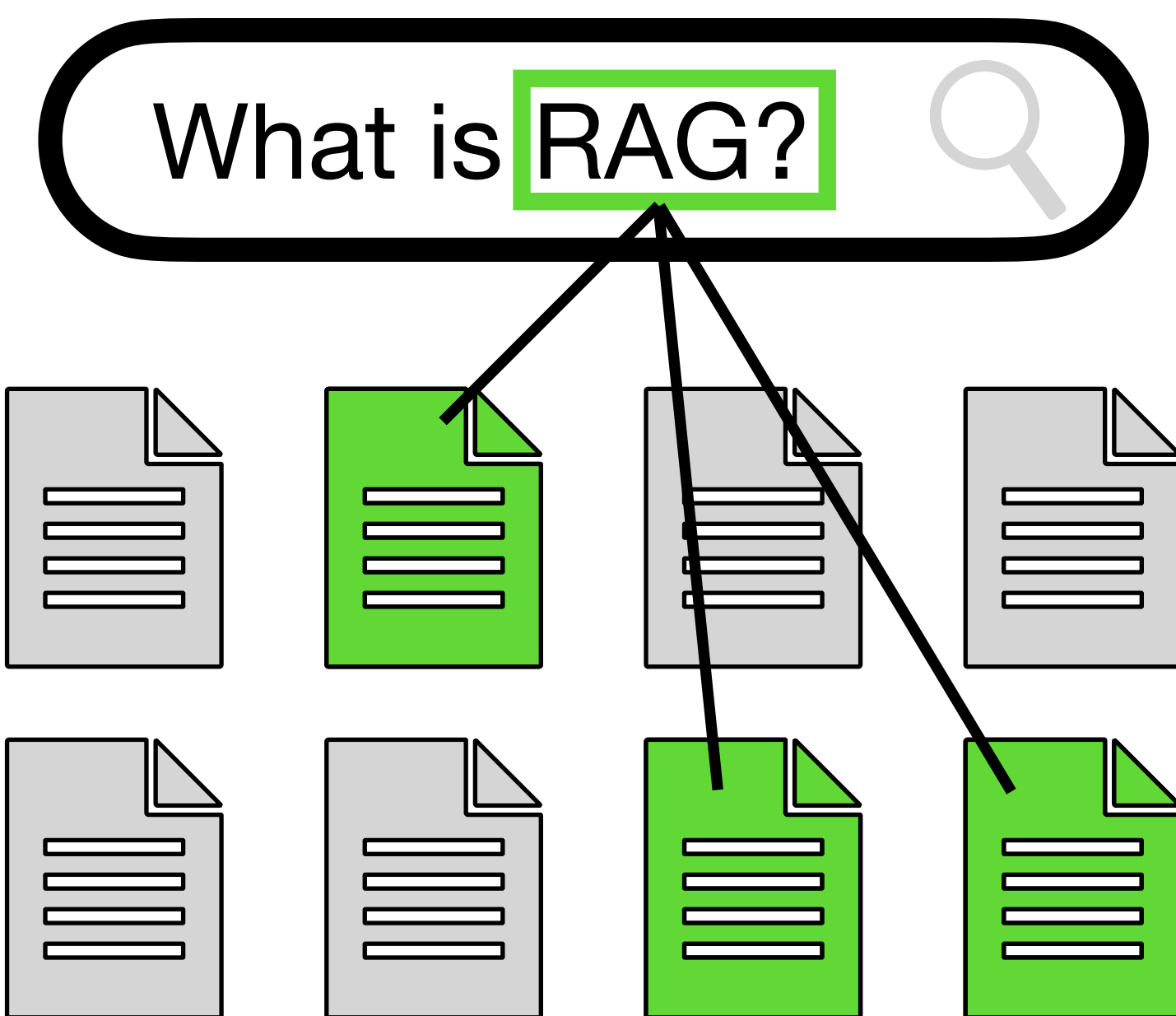


## Way 2

*Vector Search*

**How do we get these numbers?**

[−0.01, 0.02, … , 0.07]



Query          Knowledge Base

ABB #3 - Winter 2025
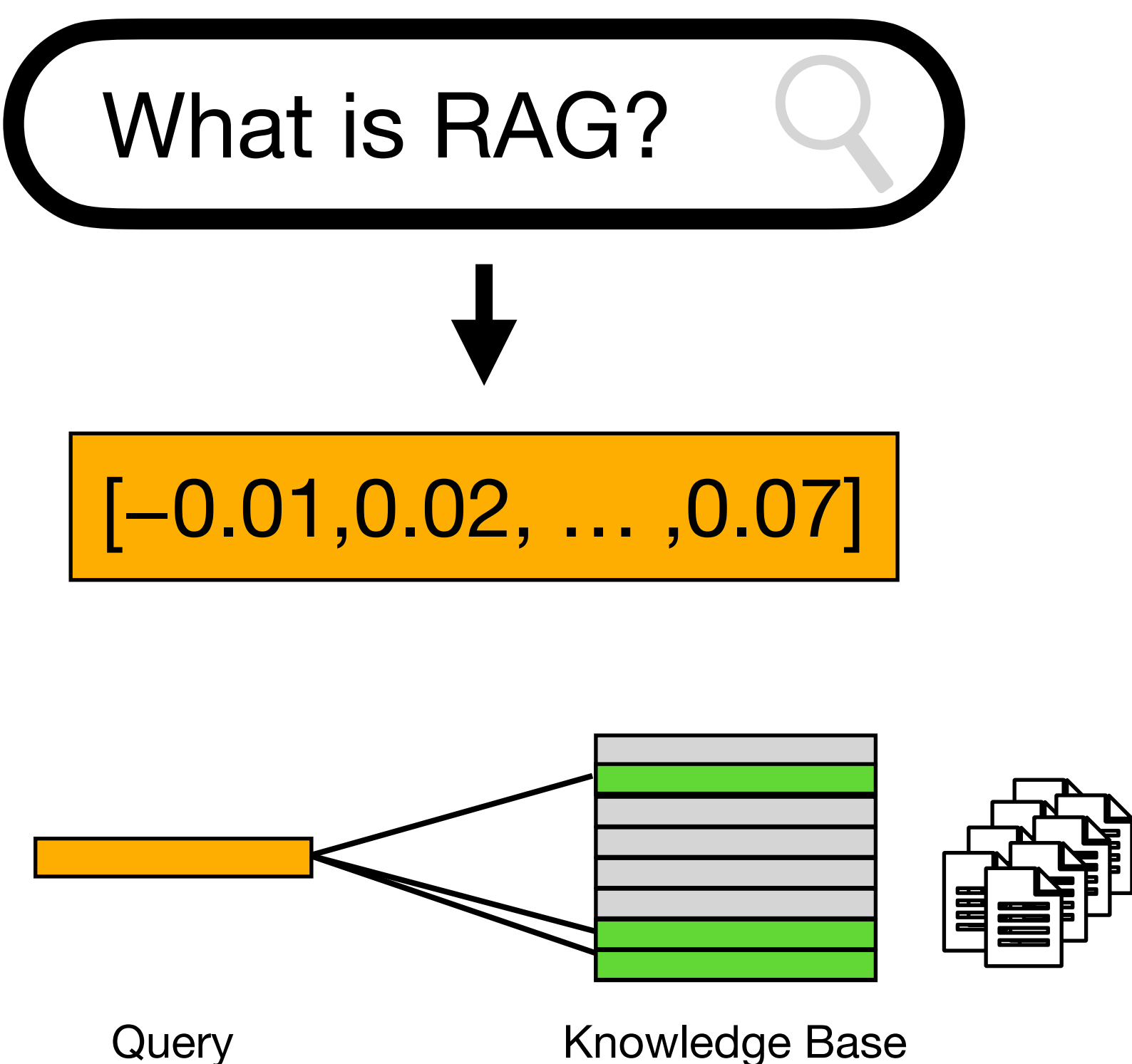
# Text Embeddings

Translate words into numbers

| Job Description | Text Embedding |
|---|---|
| Data Analyst in retail, 5 yrs | (3.4, 1.5) |
| ML Engineer in fintech, 3 yrs | (1.8, 3.1) |
| Data Scientist in healthcare, 10+ yrs. | (6.6, 2) |
| Database Admin for e-commerce, 7 yrs. | (4.2, 5) |
| BI Analyst in hospitality, early career. | (0.5, 1) |
| Data Architect, 15 yrs. | (6.5, 5) |
| Freelance Data Visualization Specialist, 4 yrs. | (2.6, 1.6) |
| Senior Data Engineer in automotive, 8 yrs. | (5, 5.5) |

ABB #3 - Winter 2025

# Text Embeddings

Translate words into *meaningful* numbers



| Job Description |
|---|
| Data Analyst in retail, 5 yrs |
| ML Engineer in fintech, 3 yrs |
| Data Scientist in healthcare, 10+ yrs. |
| Database Admin for e-commerce, 7 yrs. |
| BI Analyst in hospitality, early career. |
| Data Architect, 15 yrs. |
| Freelance Data Visualization Specialist, 4 yrs. |
| Senior Data Engineer in automotive, 8 yrs. |

(4.2, 5)  (6.5, 5)  (5, 5.5)

(1.8, 3.1)

(2.6, 1.6)  (3.4, 1.5)  (6.6, 2)

(0.5, 1)

Y

X
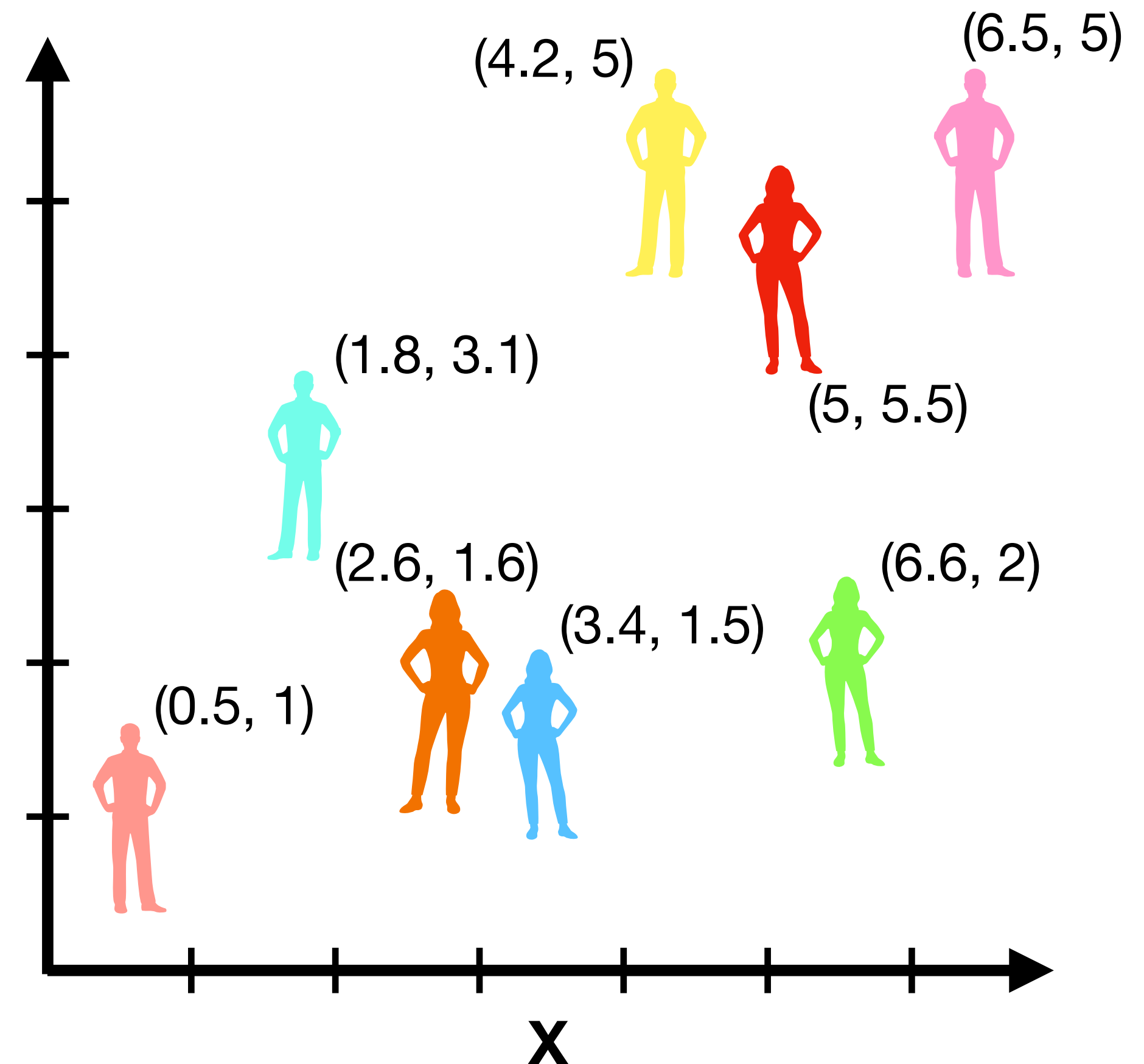
# Semantic Search

Return results based on the *meaning* of user's query



**"I need someone to build my data infrastructure"**

ABB #3 - Winter 2025
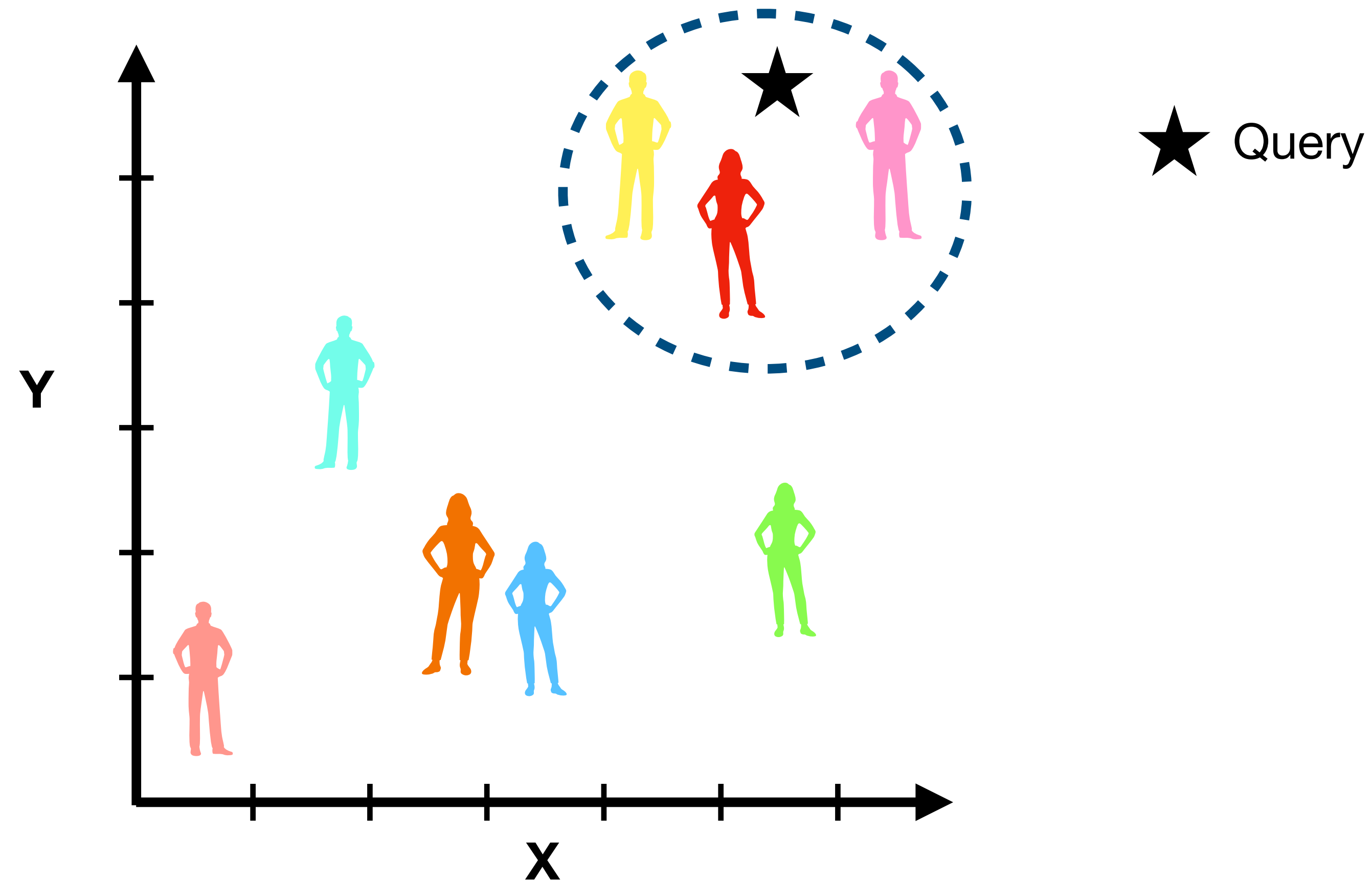
# Computing Similarities
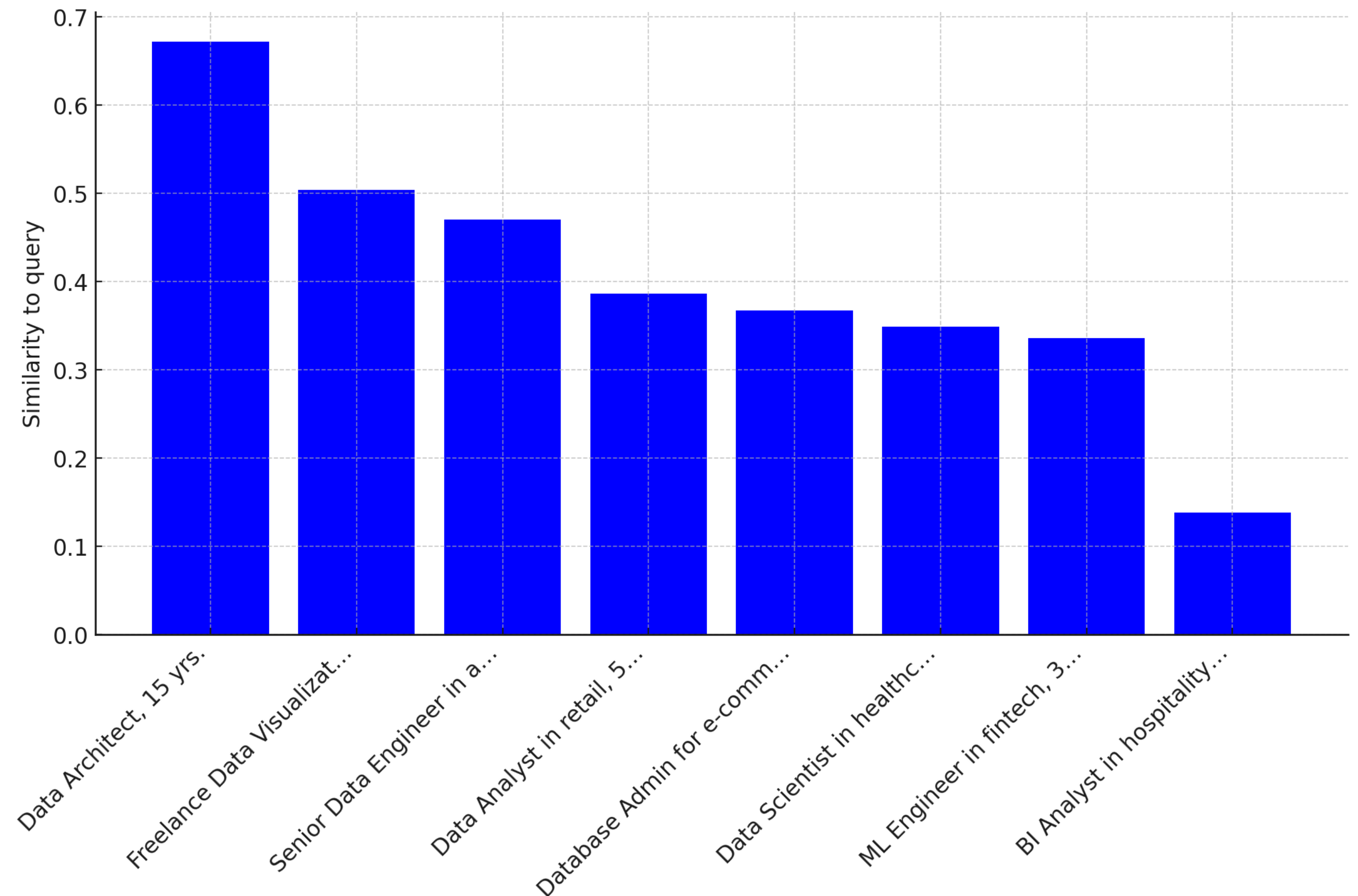
Cosine similarity

$\vec{q}$ = Query embedding

$\vec{d_i} = i^{th}$ JD embedding

**Cosine Similarity**

$$s_i = \frac{\vec{q} \cdot \vec{d_i}}{\|\vec{q}\|\|\vec{d_i}\|}$$

**Cosine Similarities between $\vec{q}$ and $\vec{d_i}$**

ABB #3 - Winter 2025

# Similarity Scores
Scaling cosine similarity with softmax

**Softmax Function**

$$\bar{s}_i = \frac{e^{\frac{s_i}{\tau}}}{\sum e^{\frac{s_i}{\tau}}}$$

$s_i$ = cosine similarity between $\vec{q}$ and $\vec{d}_i$

$\tau$ = Temperature $\quad \sum \bar{s}_i = 1$

**Probability-like score!**



**Low Temperature ($\tau = 0.1$)**

ABB #3 - Winter 2025

# Similarity Scores
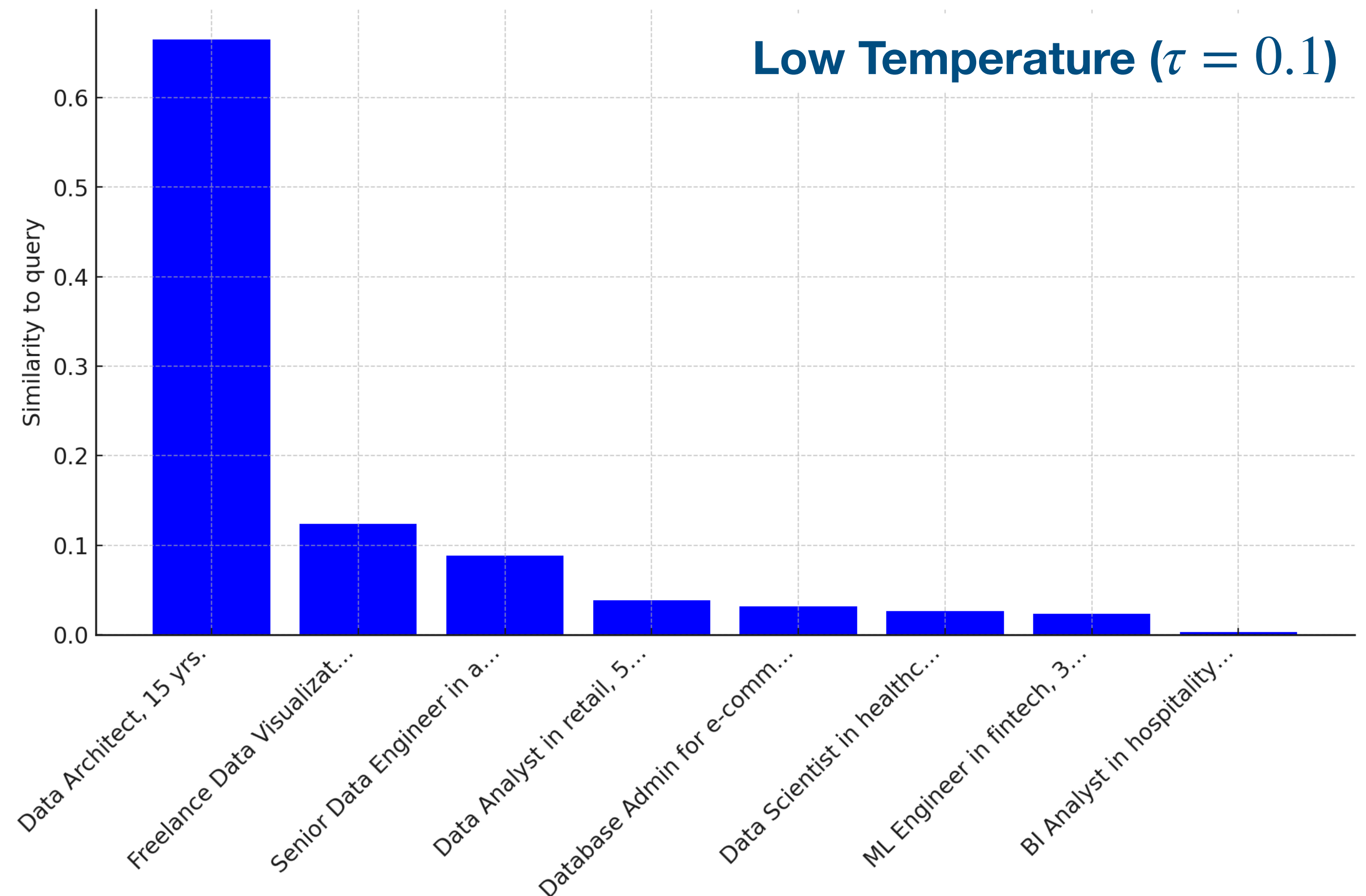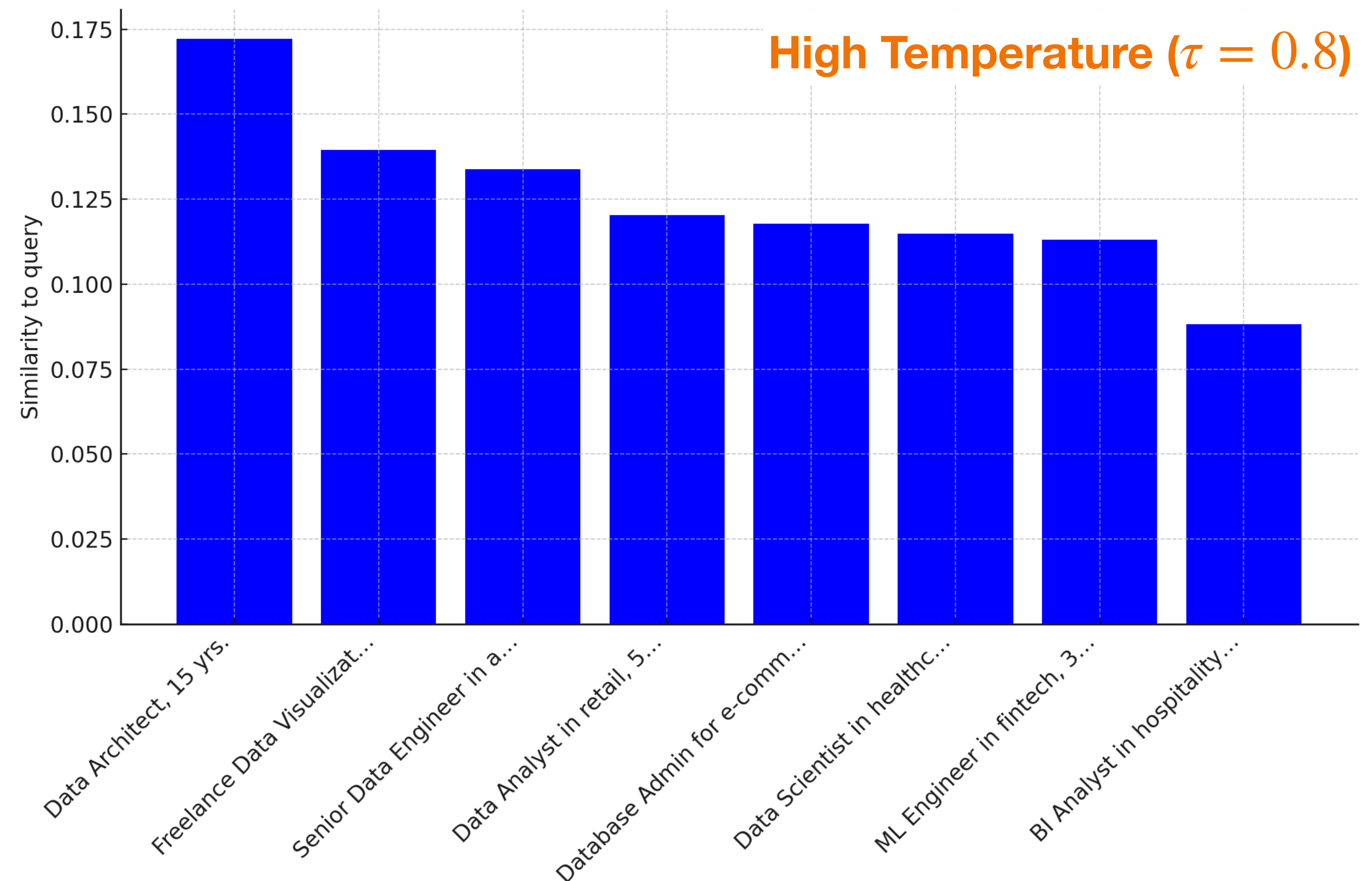## Scaling cosine similarity with softmax

**Softmax Function**

$$\bar{s}_i = \frac{e^{\frac{s_i}{\tau}}}{\sum e^{\frac{s_i}{\tau}}}$$

$s_i$ = cosine similarity between $\vec{q}$ and $\vec{d}_i$

$\tau$ = Temperature $\qquad \sum \bar{s}_i = 1$

**Probability-like score!**

**High Temperature ($\tau = 0.8$)**

ABB #3 - Winter 2025

# More Use Cases

Embeddings make text computable!
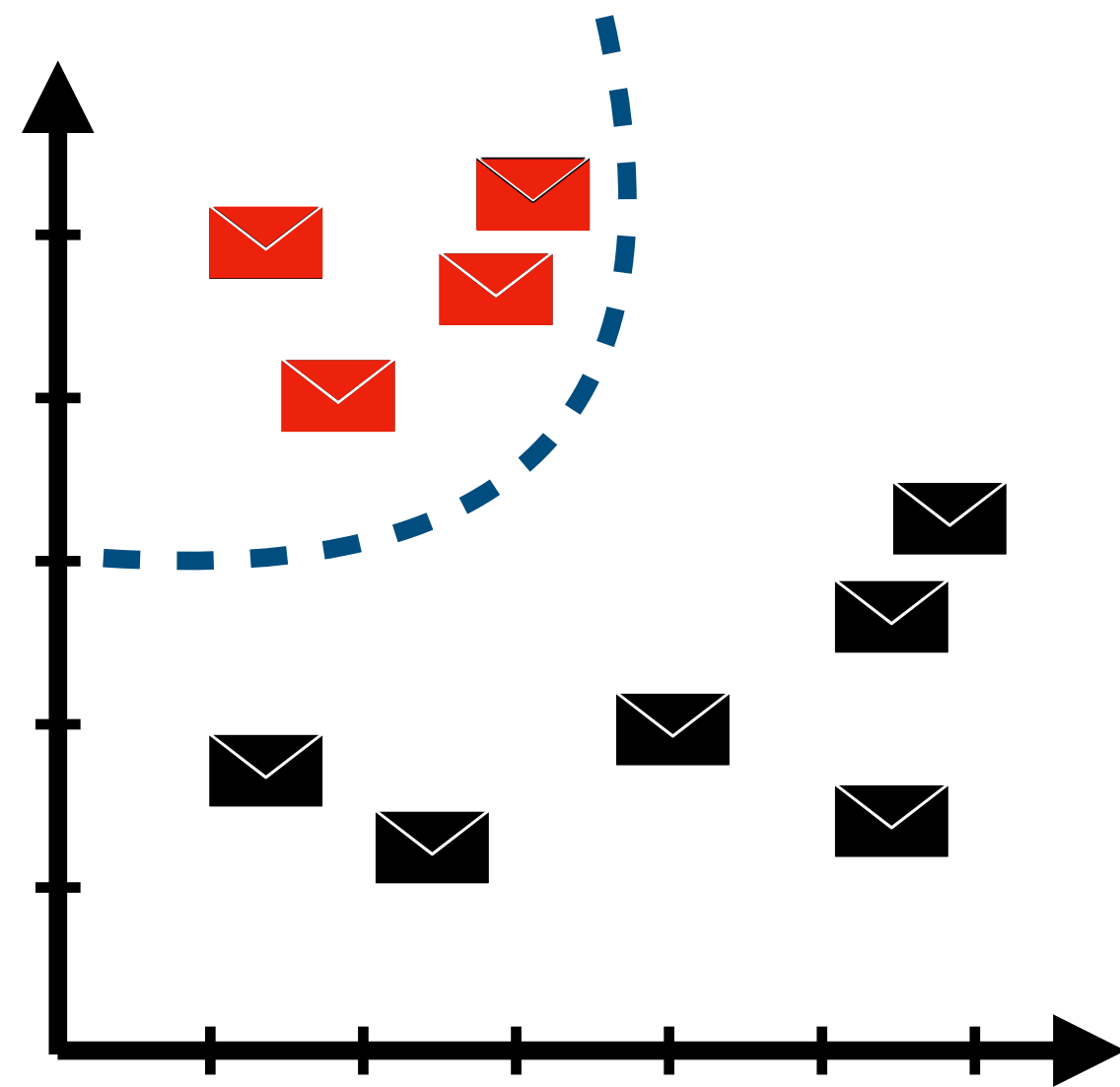


**Text Classification**

*e.g. spam detection*

**Clustering**

*e.g. ICA analysis*

**Regression**

*e.g. product sales forecasting*

ABB #3 - Winter 2025

# Examples

ABB #3 - Winter 2025

# Example 1
## Analyzing Survey Data with Embeddings (Motivation)

ABB #3 - Winter 2025

# Example 1

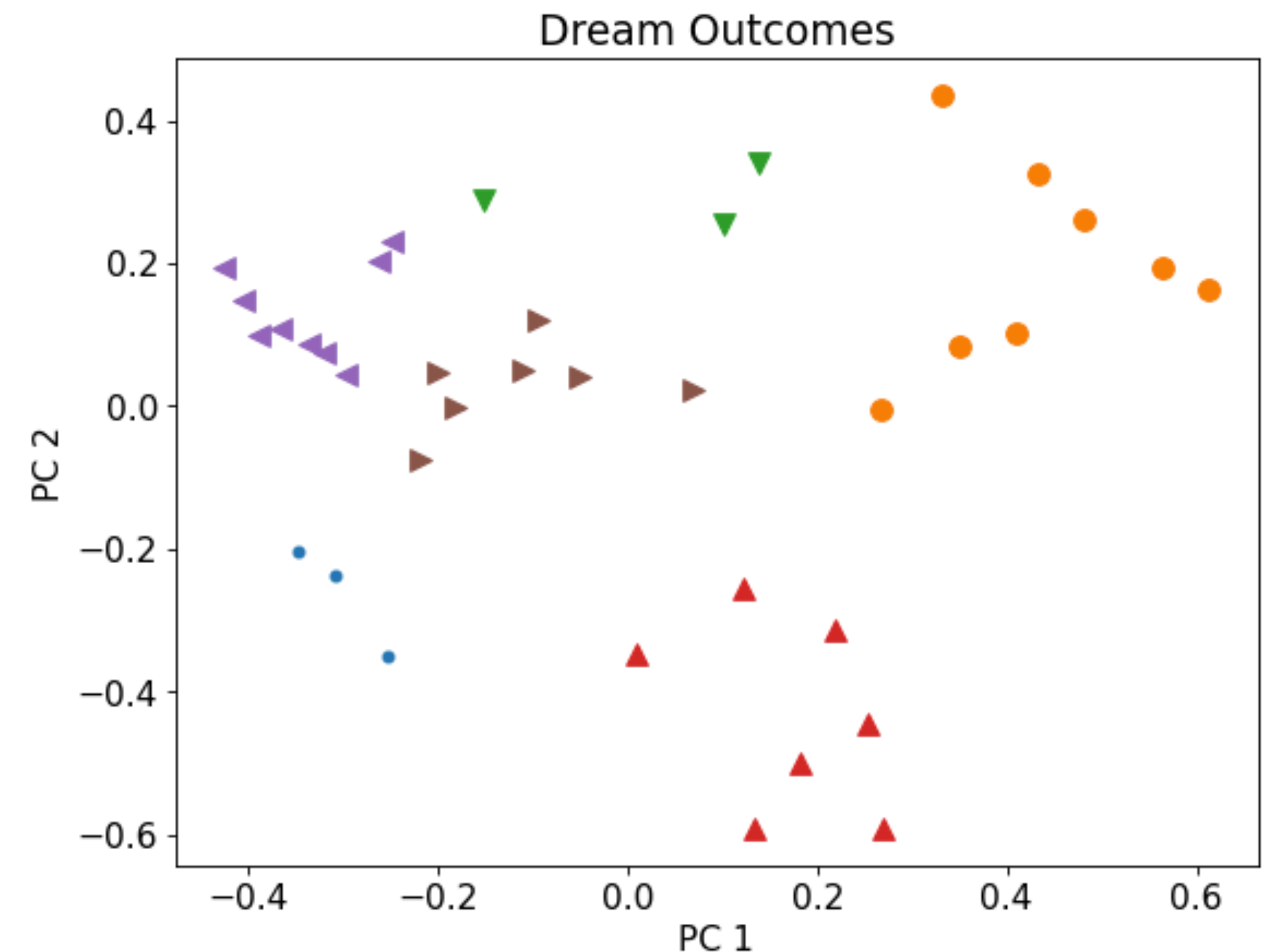## Analyzing Survey Data with Embeddings (Motivation)

- Hands on projects, and switch to Data science career path from research
- Get to know Shaw better! Plus, have a clear path with resources to guide you on where to go and what to do for the future
- Fundamentals understanding, Hands on skills, small projects running in diff frameworks, create a small GH portfolio.
- Build multiple AI applications
- I would like to learn how to practically setup LLM application in a production environment so that I can start creating some AI web apps for internal and external use.
- Ability to train model based on a data set, and how to do predictive analysis.
- Develop a good foundation in AI/ML.
- I love to figure out good strategies to use in my software solutions using AI
- Learn to decide and advice between different data science archictures and options
- Implement practical AI cases
- Foundation in AI/ML
- Land a program manager job working with AI product/program
- Build an app
- able to create a LLM from scratch
- Get a new job
- At my former job I had a coworker who had the tedious task of making a weekly report summarizing local news for the boss. At the time I was completely sure that was something AI could do, but I didn't have the time nor the knowledge to develop such a thing. I constantly find myself having ideas like this, so a dream outcome would be finishing the course with at least a roadmap for making such an app.
- Build cool products
- Be able to be comfortable with GenAi
- A certificate to showcase my new skills, A jump start to do my own programing and be able to communicate with programmers
- Setup my own environment to compare Machine Learning statistics created by my companies Data Scientists against my own environment.
- Be able to create and deploy my own Ai powered apps
- Build and launch a product
- Getting skills to build AI technologies for many projects
- Being able to implement llms into projects
- Landing a new job
- to deepen my understanding of advanced AI techniques.And I also aim to expand my professional network and collaborate with like-minded individuals to explore new opportunities in the AI landscape.
- Be able to train a model and have some practical usage of llm.
- Learn to identify the ML solution and lead projects based on AI
- Able to put into practice AI/ML for real bunsiess solutions
- Master AI powered productivity tools to streamline regulatory compliance work
- Learn Python related to AI. Implement prototypes.
- 1) build GPT that accesses functions and APIs, 2) hands-on fine-tune a model using LoRA 3) write and debug Python code (with AI assistance) that accesses ChatGPT, Perplexity and Google search, 4) build working agent(s) that can output results in less than 10 seconds 5) possibly use a RAG efficiently for booklength PDF texts
- Fully understand AI and LLM, and how to build one using Python
- Be able to comfortably add AI into daily or business use cases
- Build my own AI product with the AI services available these days
- I want to build my own MicroSaaS products
- I love to apply ai solution to existing and future back-end projects

What is your dream outcome for this course? *

Your answer

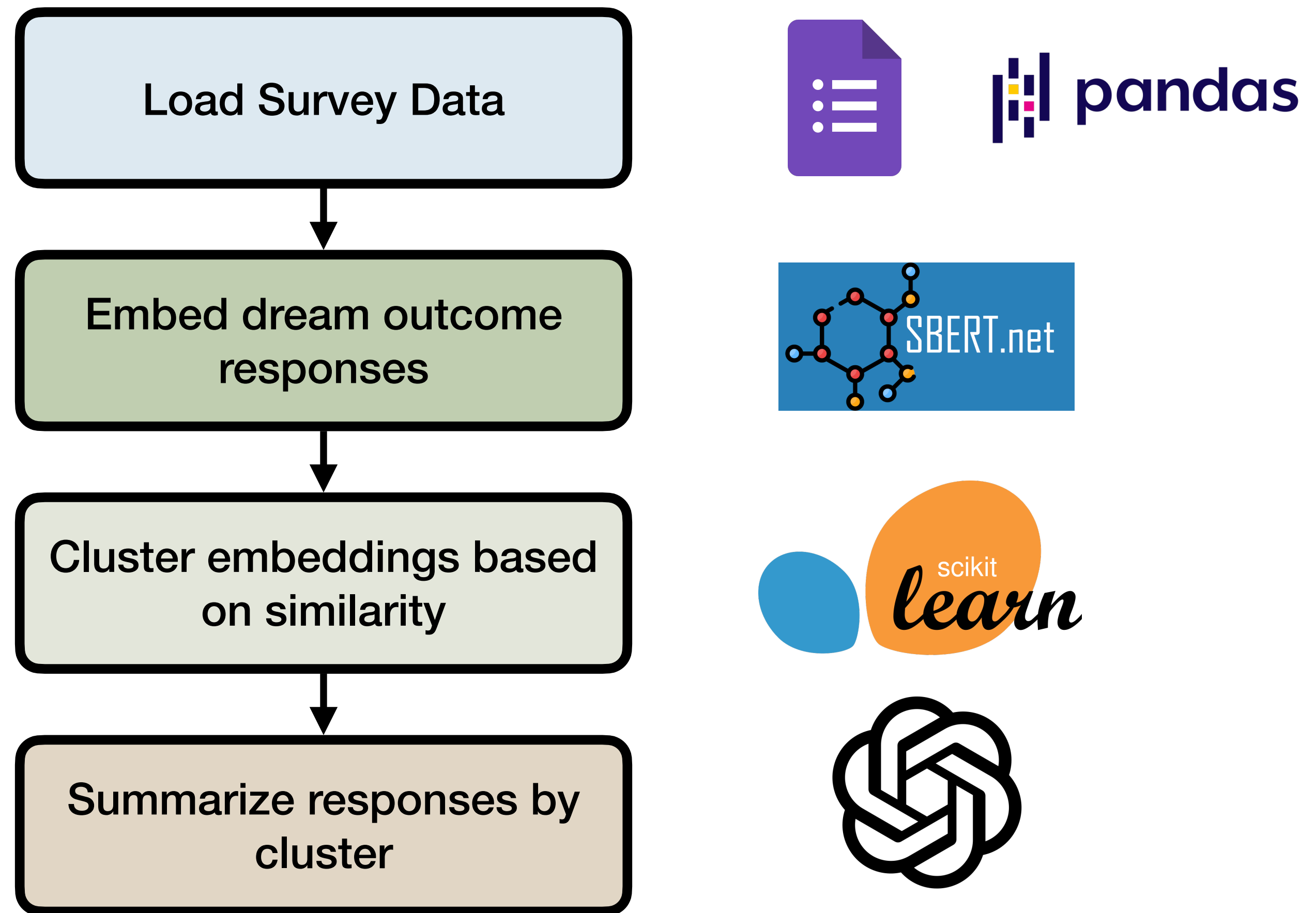Objective, quantitative analysis is **challenging.**

# Example 1
## Analyzing Survey Data with Embeddings (Overview)

- Hands on projects, and switch to Data science career path from research
- Get to know Shaw better! Plus, have a clear path with resources to guide you on where to go and what to do for the future
- Fundamentals understanding, Hands on skills, small projects running in diff frameworks, create a small GH portfolio.
- Build multiple AI applications
- I would like to learn how to practically setup LLM application in a production environment so that I can start creating some AI web apps for internal and external use.
- Ability to train model based on a data set, and how to do predictive analysis.
- Develop a good foundation in AI/ML.
- I love to figure out good strategies to use in my software solutions using AI
- Learn to decide and advice between different data science archictures and options
- Implement practical AI cases
- Foundation in AI/ML
- Land a program manager job working with AI product/program
- Build an app
- able to create a LLM from scratch
- Get a new job
- At my former job I had a coworker who had the tedious task of making a weekly report summarizing local news for the boss. At the time I was completely sure that was something AI could do, but I didn't have the time nor the knowledge to develop such a thing. I constantly find myself having ideas like this, so a dream outcome would be finishing the course with at least a roadmap for making such an app.
- Build cool products
- Be able to be comfortable with GenAi
- A certificate to showcase my new skills, A jump start to do my own programing and be able to communicate with programmers
- Setup my own environment to compare Machine Learning statistics created by my companies Data Scientists against my own environment.
- Be able to create and deploy my own Ai powered apps
- Build and launch a product
- Getting skills to build AI technologies for many projects
- Being able to implement llms into projects
- Landing a new job
- to deepen my understanding of advanced AI techniques.And I also aim to expand my professional network and collaborate with like-minded individuals to explore new opportunities in the AI landscape.
- Be able to train a model and have some practical usage of llm.
- Learn to identify the ML solution and lead projects based on AI
- Able to put into practice AI/ML for real bunsiess solutions
- Master AI powered productivity tools to streamline regulatory compliance work
- Learn Python related to AI. Implement prototypes.
- 1) build GPT that accesses functions and APIs, 2) hands-on fine-tune a model using LoRA 3) write and debug Python code (with AI assistance) that accesses ChatGPT, Perplexity and Google search, 4) build working agent(s) that can output results in less than 10 seconds 5) possibly use a RAG efficiently for booklength PDF texts
- Fully understand AI and LLM, and how to build one using Python
- Be able to comfortably add AI into daily or business use cases
- Build my own AI product with the AI services available these days
- I want to build my own MicroSaaS products
- I love to apply ai solution to existing and future back-end projects



Dream Outcomes

ABB #3 - Winter 2025

# Example 1

## Analyzing Survey Data with Embeddings (Flowchart)

ABB #3 - Winter 2025

# Example 1
Analyzing Survey Data with Embeddings (Code)

ABB #3 - Winter 2025

# Example 2
## Semantic Search with Embeddings (Motivation)



**Finding specific information across several resources can be tedious**

ABB #3 - Winter 2025

# Example 2
## Semantic Search with Embeddings (Overview)

What is RAG? 🔍

**How to Improve LLMs with RAG**

A beginner-friendly introduction w/ Python code

Shaw Talebi
Published in Towards Data Science · 13 min read · Mar 9, 2024

815    💬 6

1. **Article title:** How to Improve LLMs with RAG
   **Section:** What is RAG?
   **Snippet:** RAG works by adding a step to this basic process. Namely, a retrieval step is performed where, based on the user's prompt, the relevant information is extracted from an external knowledge base and injected into the prompt before being passed to the LLM.

2. **Article title:** How to Improve LLMs with RAG
   **Section:** What is RAG?
   **Snippet:** The basic usage of an LLM consists of giving it a prompt and getting back a response.
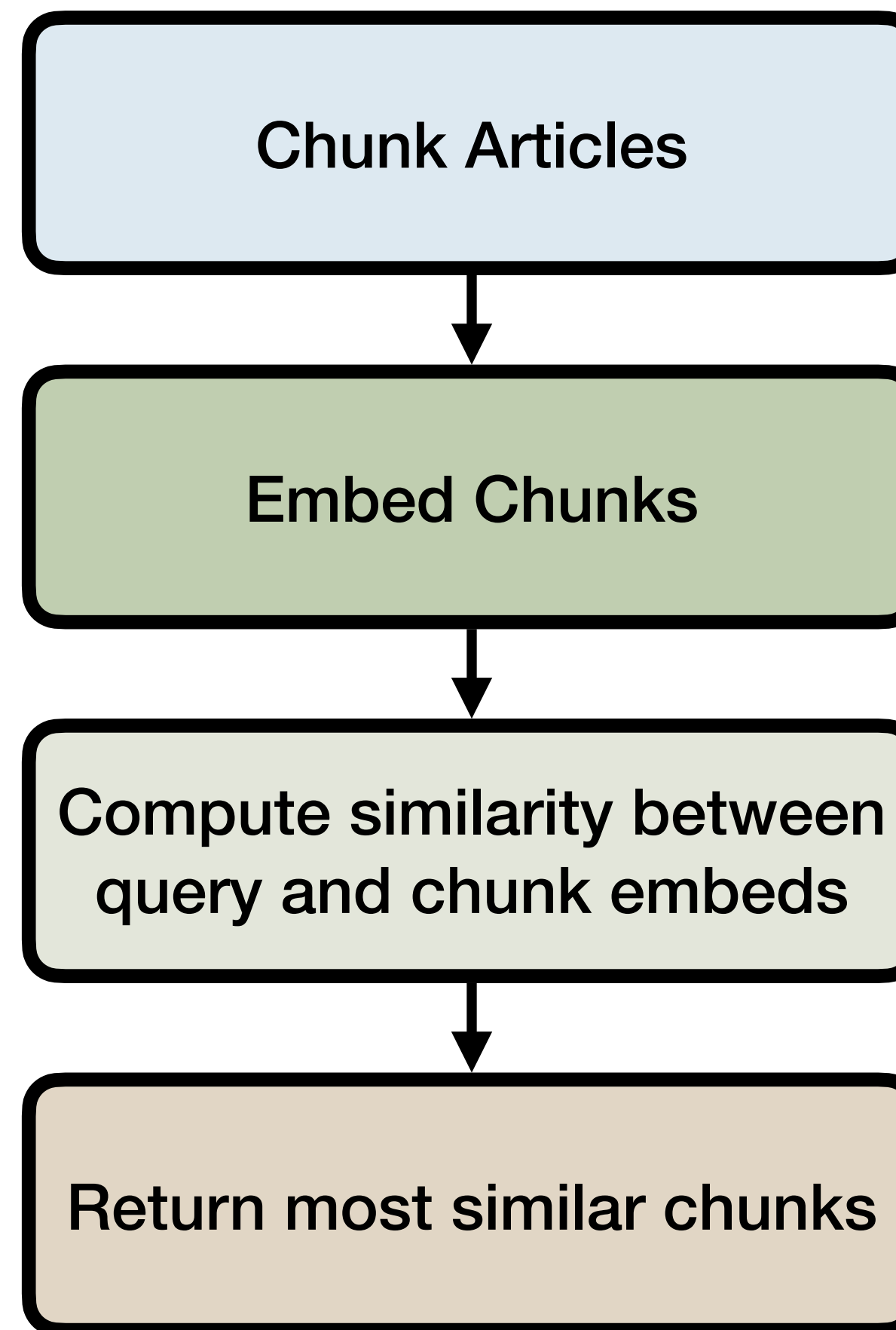
3. **Article title:** How to Improve LLMs with RAG
   **Section:** How it works
   **Snippet:** There are 2 key elements of a RAG system: a retriever and a knowledge base.

ABB #3 - Winter 2025

# Example 2
## Semantic Search with Embeddings (Flowchart)

ABB #3 - Winter 2025

# Example 2
Semantic Search with Embeddings (Code)

ABB #3 - Winter 2025

# Example 3
## Blog QA Assistant with RAG (Motivation)

What is RAG? 🔍

## How to Improve LLMs with RAG

A beginner-friendly introduction w/ Python code

Shaw Talebi
Published in Towards Data Science · 13 min read · Mar 9, 2024

👏 815    💬 6

1. **Article title:** How to Improve LLMs with RAG
   **Section:** What is RAG?
   **Snippet:** RAG works by adding a step to this basic process. Namely, a retrieval step is performed where, based on the user's prompt, the relevant information is extracted from an external knowledge base and injected into the prompt before being passed to the LLM.

2. **Article title:** How to Improve LLMs with RAG
   **Section:** What is RAG?
   **Snippet:** The basic usage of an LLM consists of giving it a prompt and getting back a response.

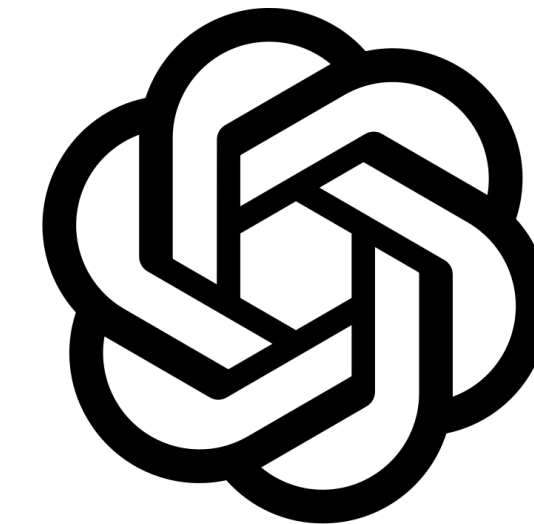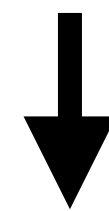3. **Article title:** How to Improve LLMs with RAG
   **Section:** How it works
   **Snippet:** There are 2 key elements of a RAG system: a retriever and a knowledge base.

**Requires digging to answer question**

ABB #3 - Winter 2025

# Example 3
## Blog QA Assistant with RAG (Overview)

What is RAG? 🔍



1. **Article title:** How to Improve LLMs with RAG
   **Section:** What is RAG?
   **Snippet:** RAG works by adding a step to this basic process. Namely, a retrieval step is performed where, based on the user's prompt, the relevant information is extracted from an external knowledge base and injected into the prompt before being passed to the LLM.

2. **Article title:** How to Improve LLMs with RAG
   **Section:** What is RAG?
   **Snippet:** The basic usage of an LLM consists of giving it a prompt and getting back a response.

3. **Article title:** How to Improve LLMs with RAG
   **Section:** How it works
   **Snippet:** There are 2 key elements of a RAG system: a retriever and a knowledge base.

RAG, or Retrieval Augmented Generation, is an approach that enhances the functionality of large language models (LLMs) by incorporating a retrieval step. This process involves extracting relevant information from an external knowledge base based on the user's prompt and injecting it into the prompt before passing it to the LLM. Essentially, while the basic operation of an LLM remains the same—receiving a prompt and generating a response—RAG augments this process by utilizing additional information to improve the quality and relevance of the output (Snippets 1, 2, 9).

A RAG system consists of two key components: a retriever, which fetches relevant information, and a knowledge base, which stores the information that can be accessed by the LLM (Snippets 3, 7). The effectiveness of a RAG system is influenced by how well the source documents are prepared, as poorly formatted documents can complicate the extraction of useful information (Snippet 5).

# Example 3
## Blog QA Assistant with RAG (Flowchart)

ABB #3 - Winter 2025
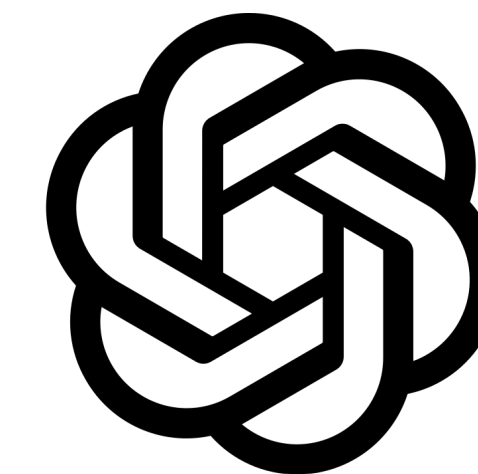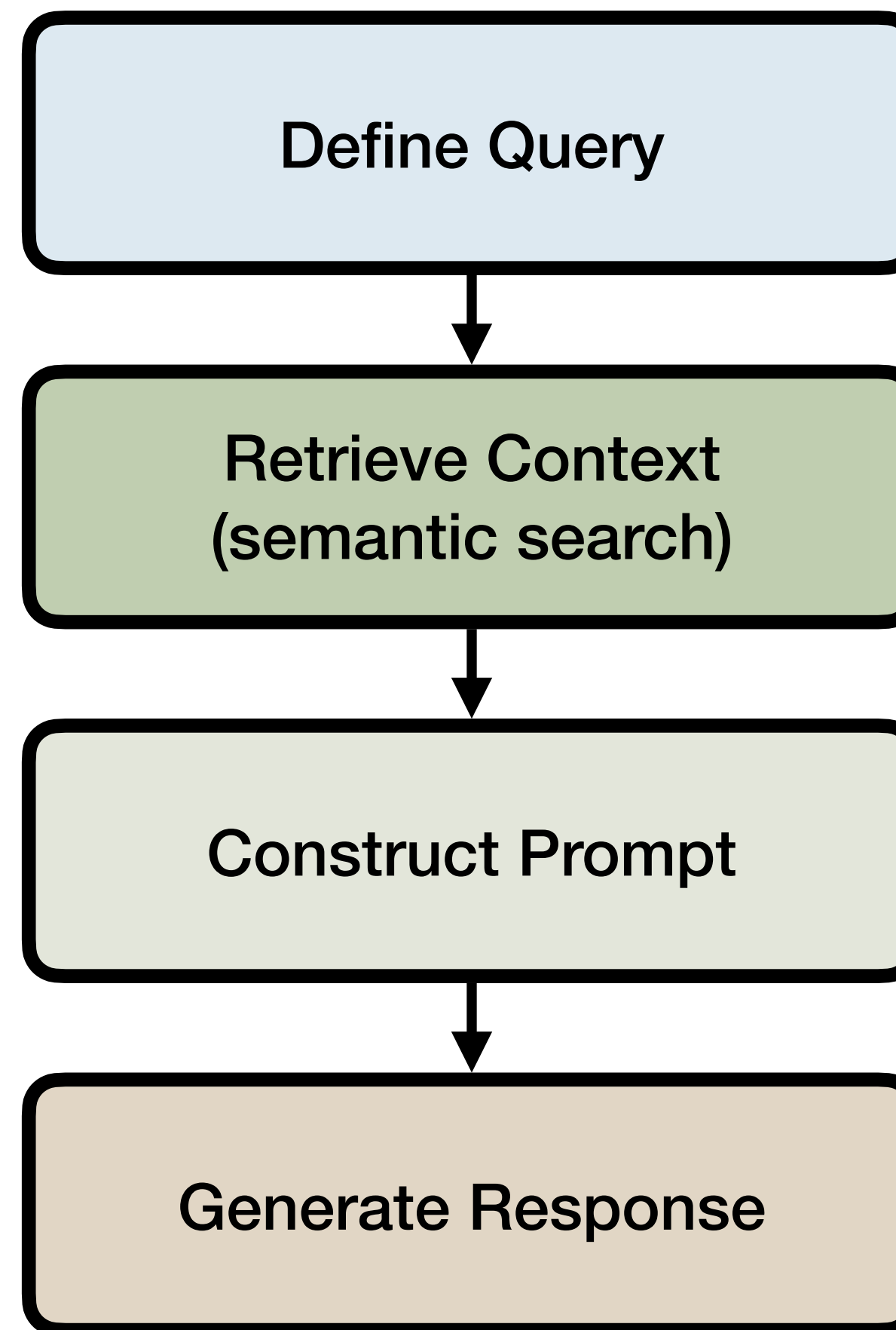
# Example 3
Blog QA Assistant with RAG (Code)

ABB #3 - Winter 2025

# Homework 4

**Project** 💻

Solve a Problem with Text Embeddings

**Pre-work** ✍️

Session 5: Fine-tuning LLMs

ABB #3 - Winter 2025

# References

[1] Multimodal RAG: Process Any File Type with AI

[2] How to Improve LLMs with RAG

[3] Text Embeddings, Classification, and Semantic Search

[4] Text Embeddings, Classification, and Semantic Search (w/ Python Code)

ABB #3 - Winter 2025

ABB #3 - Winter 2025