hack.diversity

Learning Lab: “Hack-GPT” part 2

Retrieval Augmented Generation

April 18, 2024

William Szostak

Slack: @William Szostak

LinkedIn: linkedin.com/in/william-szostak/

Email: williamszostak@gmail.com

AGENDA

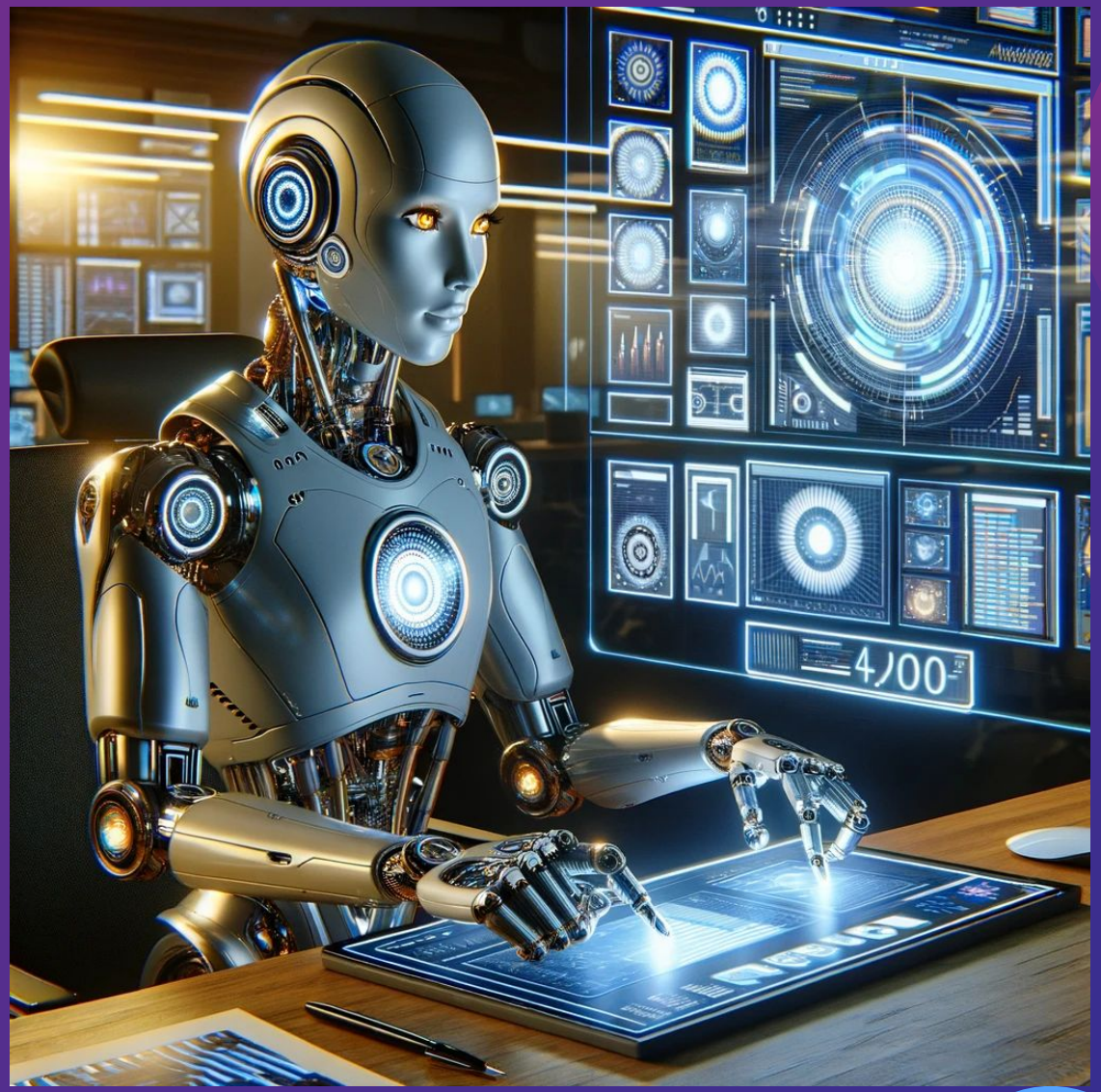
5 mins	**Intros**
10 mins	**Retrieval-Augmented What?**
10 mins	**Embeddings are the Key**
20 mins	**RAG in Action**
5 mins	**RAG vs. Fine-Tuning**
5 mins	**Wrap Up & Next Steps**

Questions welcome throughout!

hack.diversity

Disclosure:

Generative AI was used in producing these materials



About William

Professional

Education

- Went to 3 different undergraduate colleges
- Took the “10 year plan”
- BA in Mathematics

Career

- Software Engineer – Bloomberg – NYC – 1997
- Consultant - Concord, NH
- Senior Software Engineer – Portsmouth, NH
- Principal Software Engineer - Boston, MA
- Software Architect – Liberty Mutual – Portsmouth, NH - Current

Specialties

- Data Architecture
- Analytics
- Cloud Data Engineering

Passions

- Diversity, Equity, Inclusion
- Collaborative problem solving
- Innovation

Fun fact: I worked on Y2K bug fixes in 1999

Personal

I live in Rollinsford, NH
with my wife, **Phoenix** Mayet.

I enjoy kayaking,
playing sax, reading,
cooking vegetarian meals,
poker, darts, and backgammon.

We have a cat, **Rosy**,
and a dog, **Marco**.

- Favorite novel: ***Invisible Cities*** - Italo Calvino
- Favorite 80s movie: **The Blues Brothers**

Fun fact: I went to Vermont last week to see the total eclipse

... About You!

What sparks your interest in generative AI?

What are you hoping to learn today?

Retrieval-Augmented What?

hack.diversity

Retrieval Augmented Generation:

A technique for enhancing the accuracy of generative AI models with facts fetched from external sources.

nvidia.com

LLMs Have Trouble Answering Some Questions

Recent Information

All large language models were trained on information up to a certain date.

A model can't answer questions about anything that happened after that date.

Private Information

All large language models were trained on publicly available information.

A model can't answer questions about things that it wasn't trained on.

Ambiguous Terminology

Some words and phrases have different meanings in different contexts.

A model may answer questions incorrectly due to misinterpretation.

What Do LLMs Do When They Don't Know the Answer?

Hallucinations

Large language models are designed to generate responses, and in the absence of information they often make things up.

Models will confidently present false answers as if they are real.

Lack of Citations

Large language models can't provide links or references to the source of their response when there is no authoritative source.

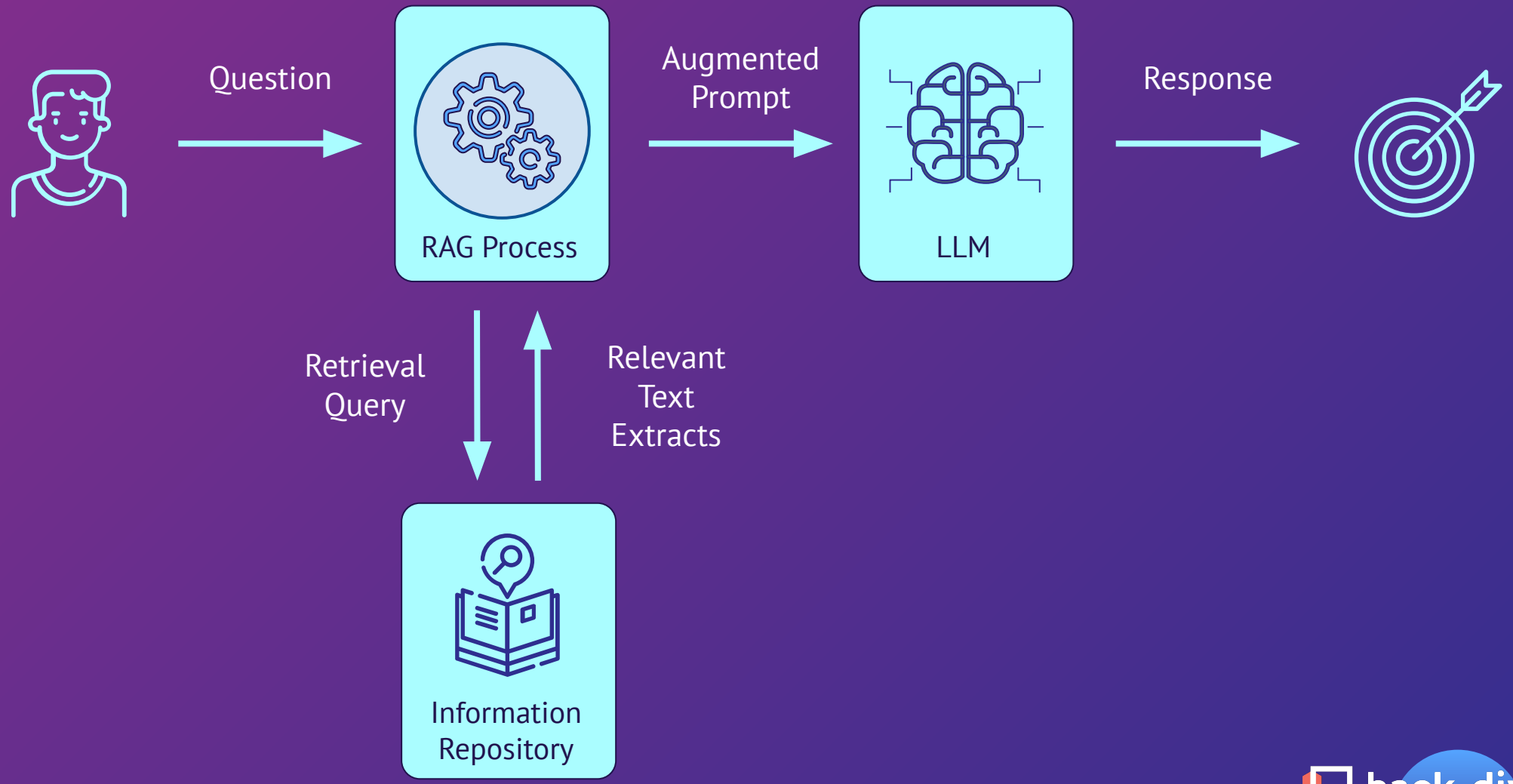
This makes is difficult to verify the accuracy of the response.

RAG to the Rescue!

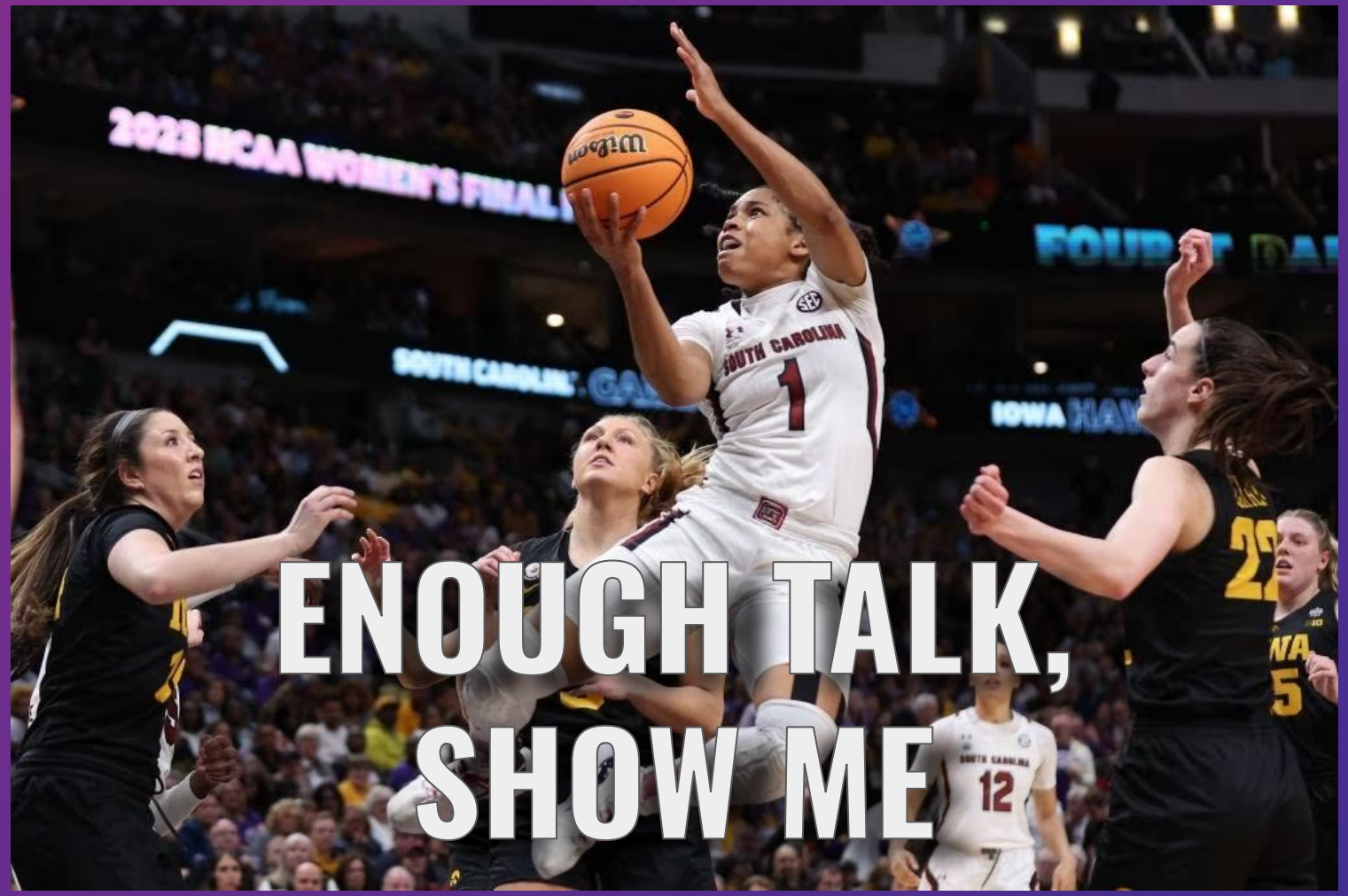
Retrieval Augmented Generation can help mitigate these issues by providing:

- Additional information
- Context
- Sources to link to

RAG: How it Works



hack.diversity



Quick Demo: **RAG at Work**

Embeddings are the Key

hack.diversity

Retrieval

We have a collection of text that we want to use to help the LLM answer questions.

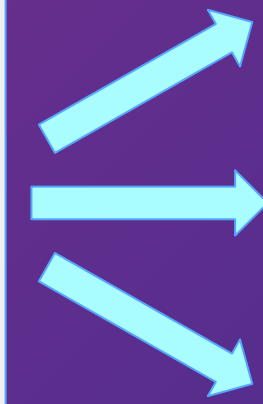
How do we find the most relevant text for a given question?

Key Word Search?

Employee Question:

How much **vacation** time do I get?

SUBMIT



Employee Handbook:

Employees receive 15 days of **Personal Time Off** (PTO) in their first year

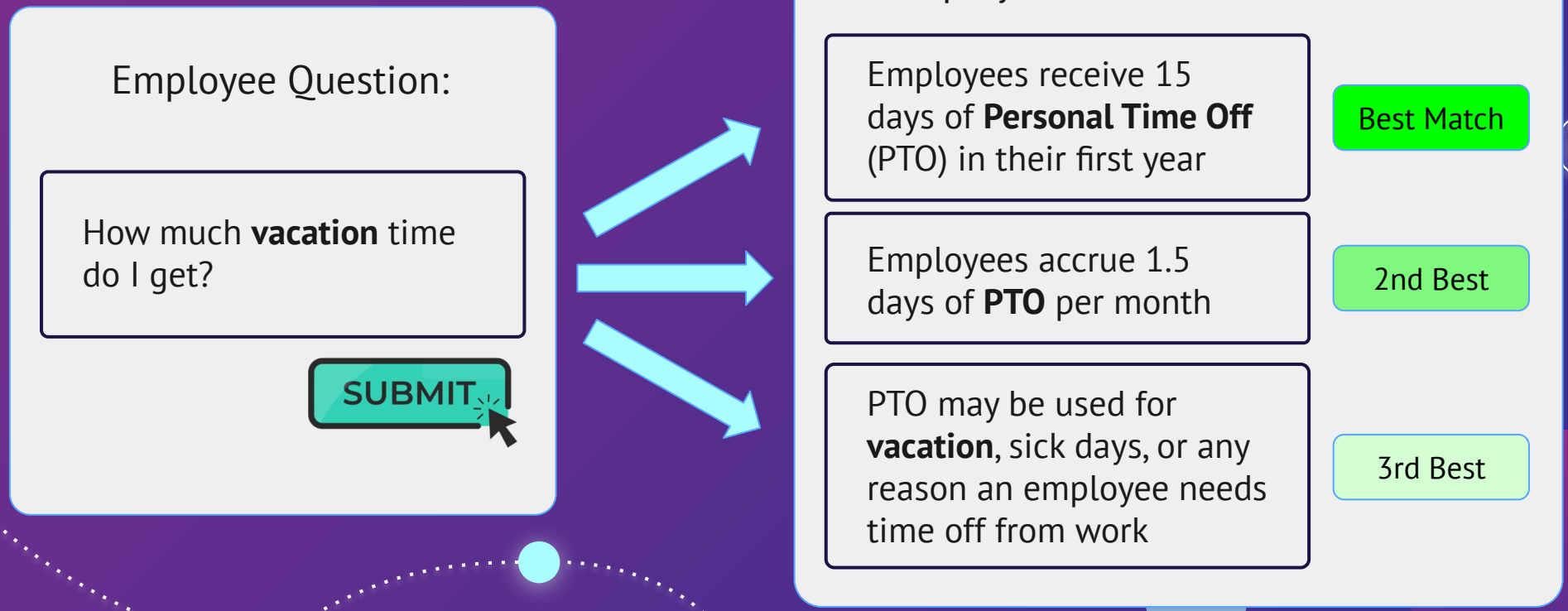
Employees accrue 1.5 days of **PTO** per month

PTO may be used for **vacation**, sick days, or any reason an employee needs time off from work

Semantic Search

Captures the **meaning** of words and their **relationships** to each other

Ranks similarity of words and phrases to return the best matches



Q: How do we represent the meaning of pieces of text in a way that allows us to compare them and rank their similarity?

A: Embeddings

Embeddings Are Vectors that Represent the Meaning of Text

Input (Text)

We the people, in order to form a more perfect union...

When you lookin' at me, tell me what do you see? (I love myself)

You must be the change you wish to see in the world.

Embedding Model

Output (Vector)

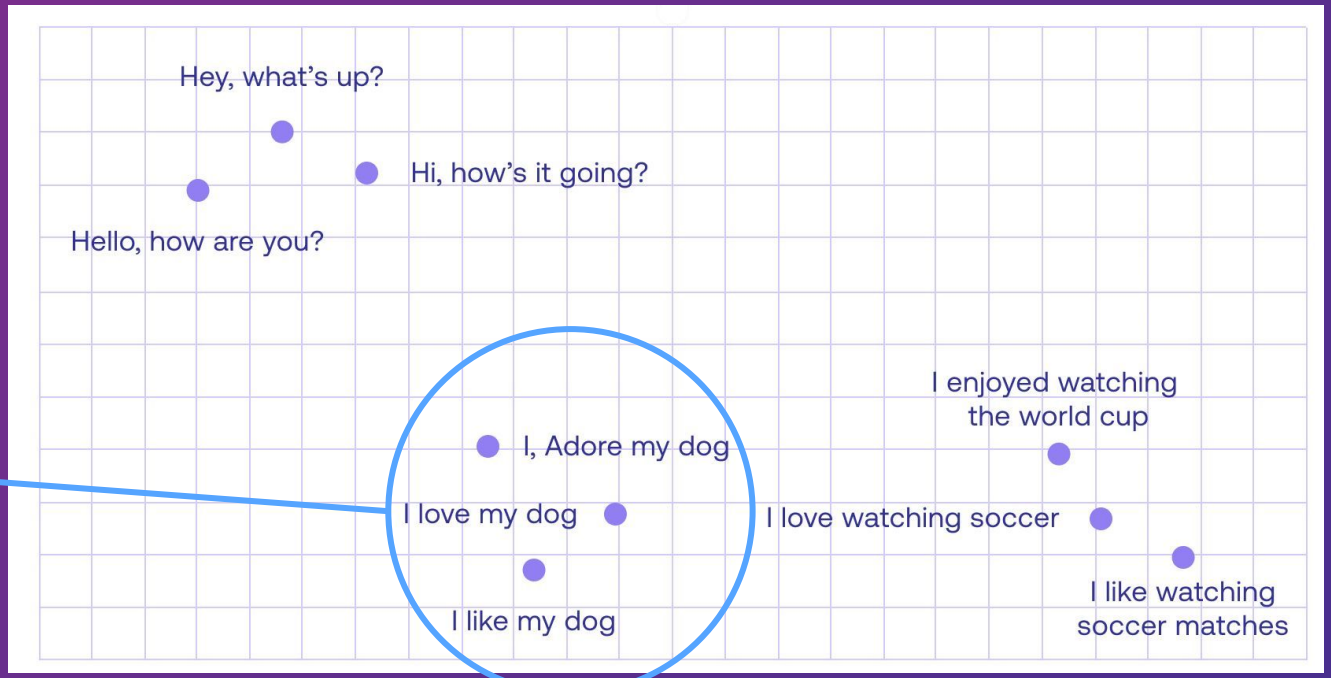
[0.0034668462, 0.017092062, 0.017092062, 0.017092062, ...]

[-0.007647873, -0.053987905, -0.0020174133, 0.044038847, ...]

[0.012555103, -0.029928306, -0.029800843, 0.059703656, ...]

A Vector Is a Set of Coordinates

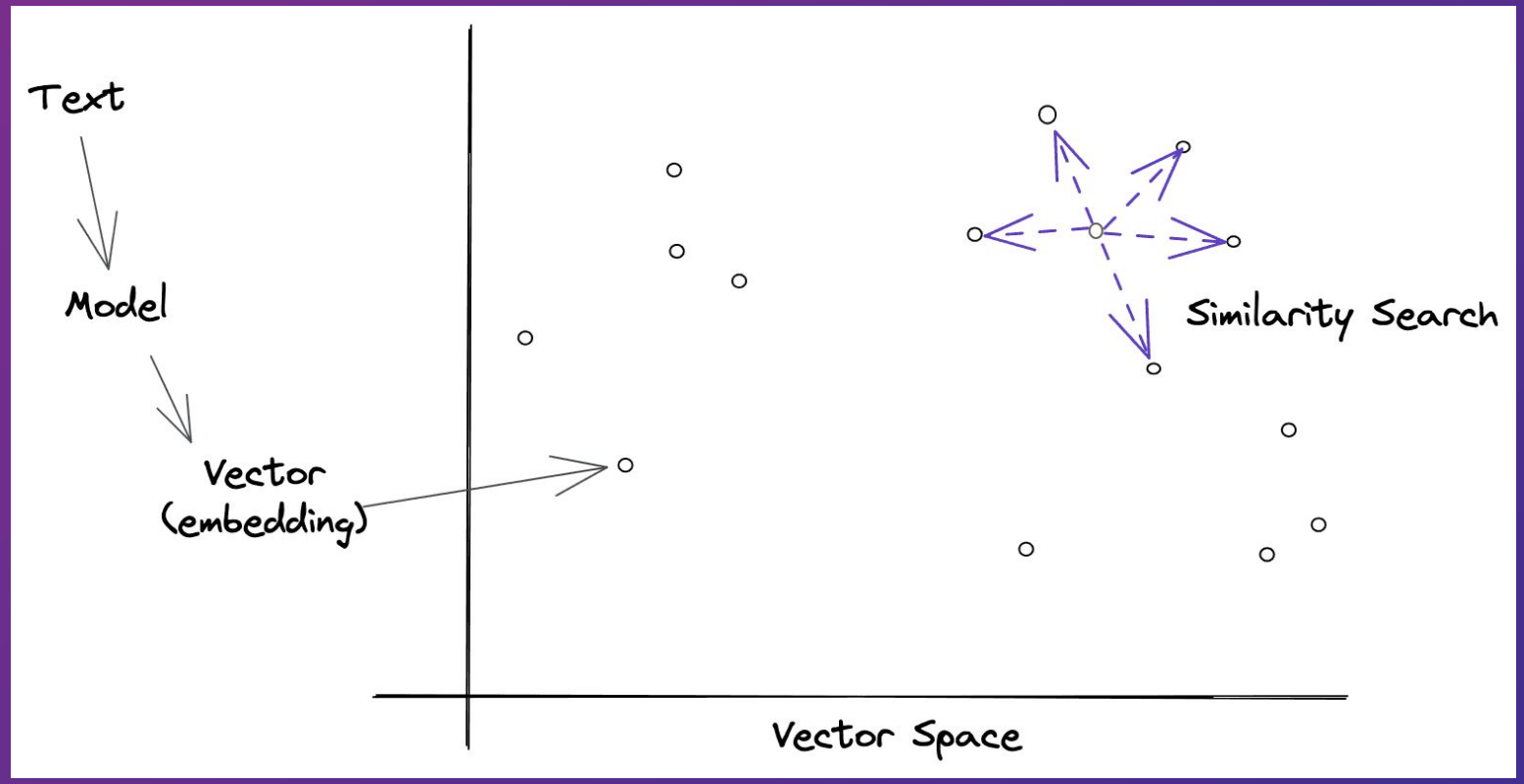
Vectors that are close to each other represent words or phrases that are similar or related



https://txt.cohere.com/sentence-word-embeddings/

I like to think of embedding vectors as representing:
Where in the model's "brain" does this concept reside?

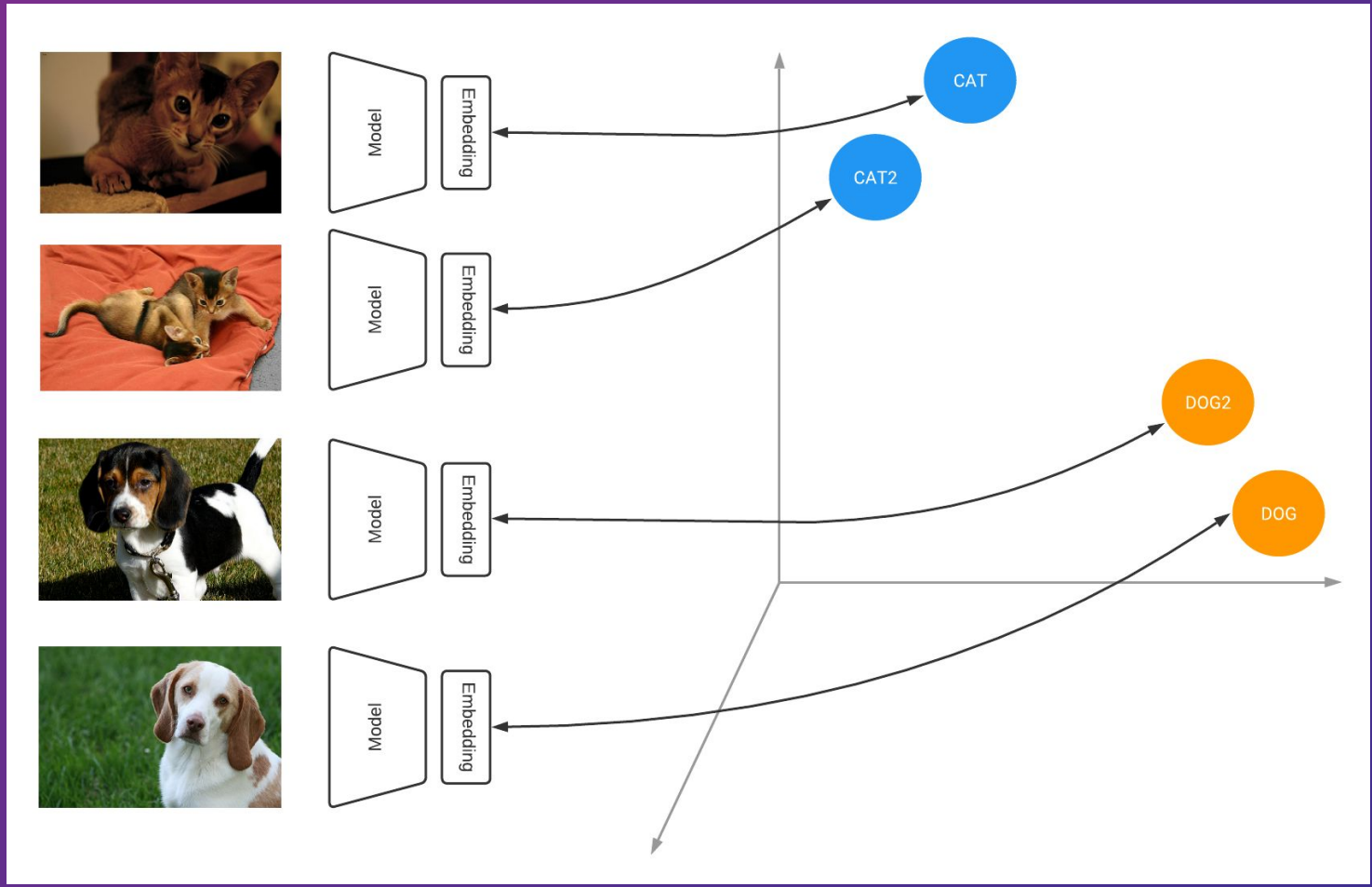
Vectors Enable Similarity Search Based on Meaning



https://www.tylercrosse.com/ideas/semantic-search

Good TO KNOW

Embeddings Can Represent Images or Audio, Too



https://blog.tensorflow.org/2021/09/introducing-tensorflow-similarity.html

Demo: **Embeddings**

RAG in Action

hack.diversity

Putting It Together

Gather Relevant Text

Chunk it up into meaningful segments

Create Embeddings

Store the embedding vectors with the text segments they represent

User Asks a Question

Create embedding vector of the user's question

Find Related Text Segments

Compare vectors to find the closest matches

Create Prompt for LLM

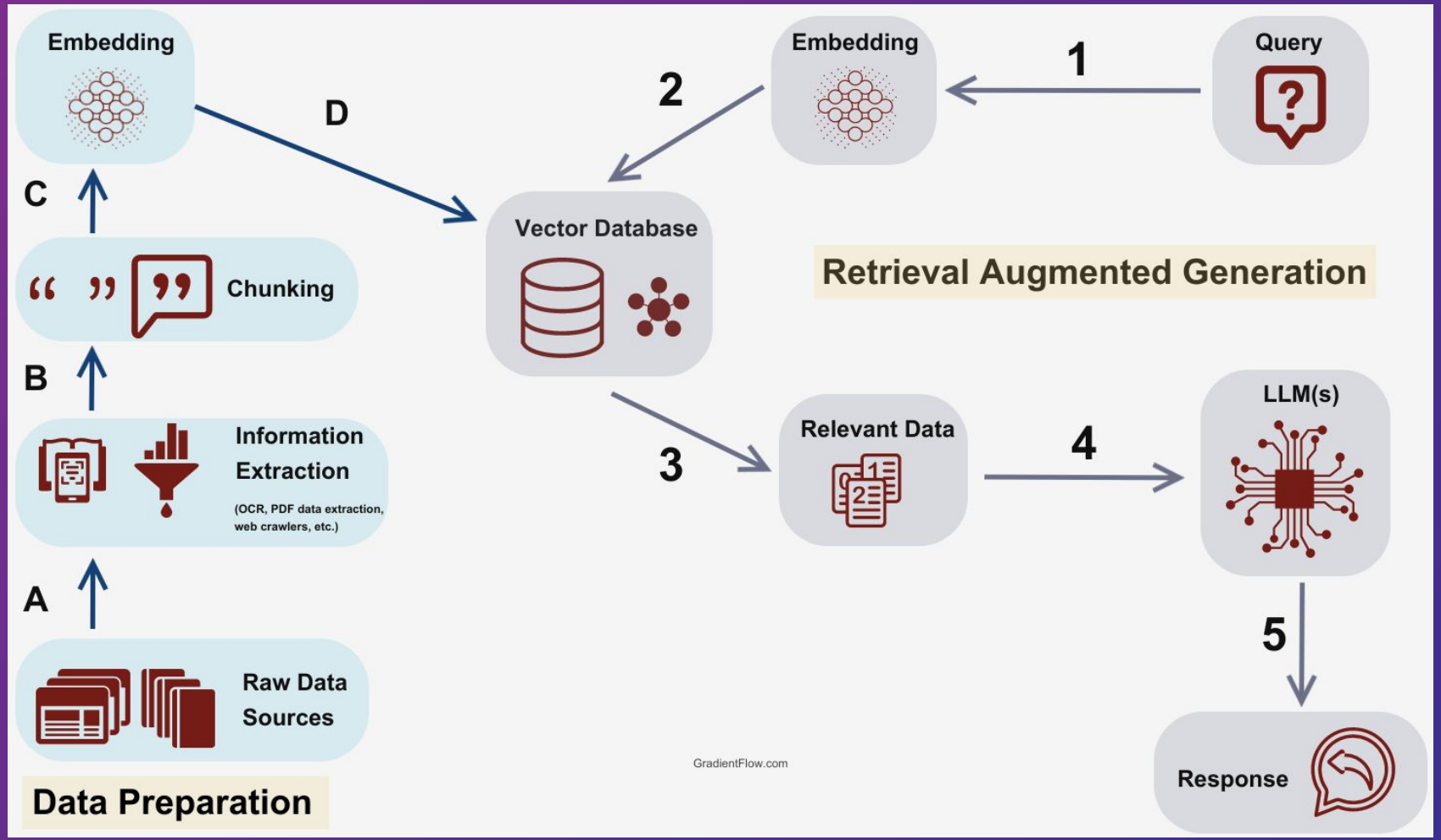
"Using the following information:
[text segments]

Answer this question:
[user question]"

LLM Answers the Question

Return the answer to the user.
Include links to the original documents.

How It Works



gradientflow.substack.com

Example Use Cases

Amazon

Source: Customer reviews

Prompt: Summarize what customers liked and didn't like about this product

Microsoft Copilot

Source: Email

Prompt: What action items do I have from my emails in the past week?

Insurance Underwriting

Source: SEC filings, news stories, web reviews

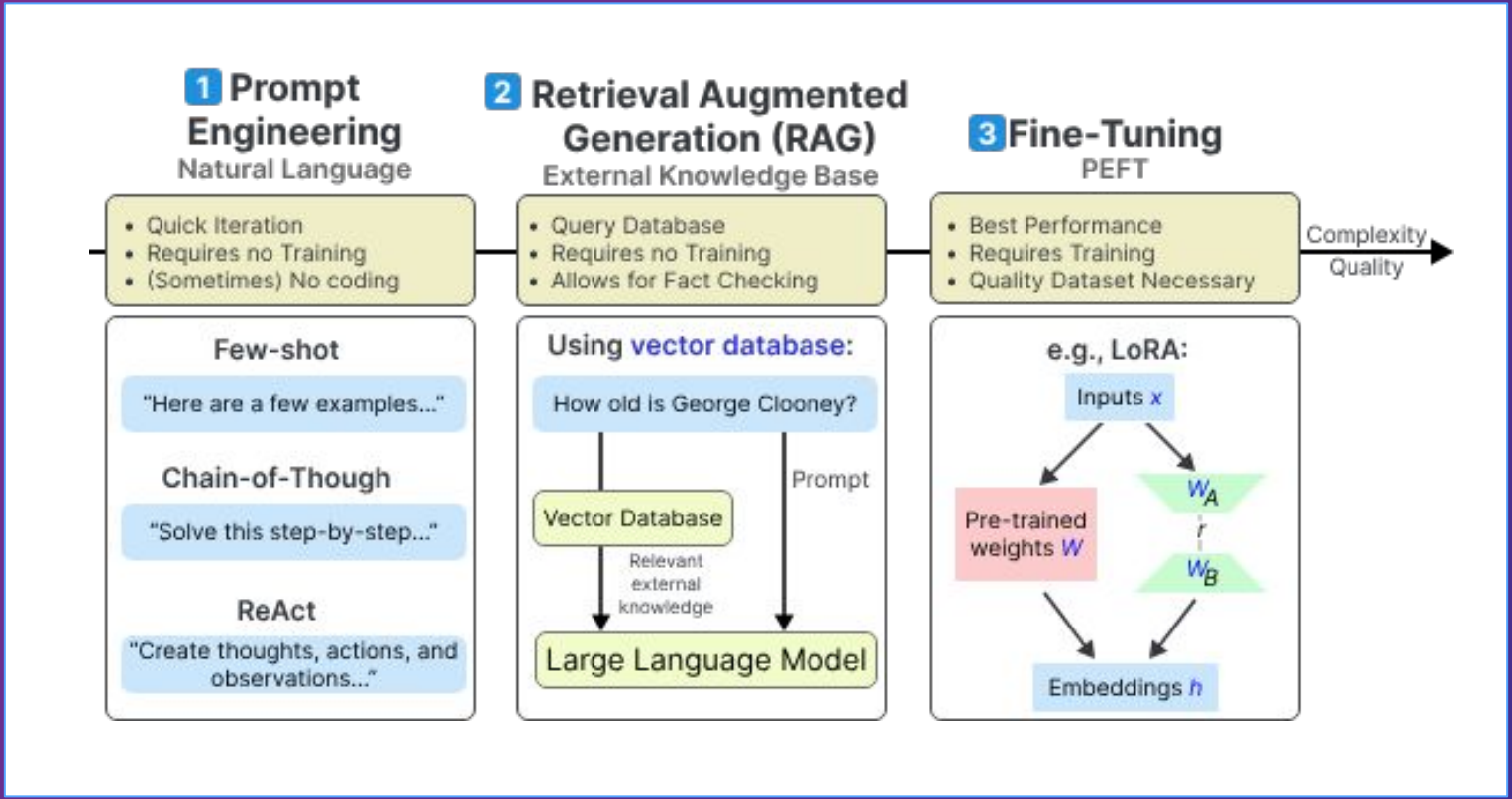
Prompt: Does this business have any of the following risk factors…

Demo: RAG in Depth

RAG vs. Fine-Tuning

hack.diversity

Complexity Comparison



https://www.maartengrootendorst.com/blog/improving-llms/

RAG is more complex than prompt engineering, but not as complex and expensive as fine-tuning.

RAG vs. Fine-Tuning: Trade-Offs

	RAG	Finetuning
External Data needed	✓	✗
Frequently changing data	✓	✗
Model behavior change (Domain knowledge)	✗	✓
Traceability	✓	✗
Training Data available	✗	✓
Minimize Hallucinations	✓	✗

Fine-tuning can give the model a more in-depth “understanding” of a new domain.

Because of the traceability, reduction in hallucinations, and ease of implementation, RAG is often a better choice.

Wrap Up & Next Steps

hack.diversity

Thank you for participating!

Please take this 3-question survey
to help me make future sessions better!

Your feedback is a gift!

https://forms.gle/QwSC5PKXtxx631Pg8

hack.diversity

Resources

Repo: Demo code: github.com/williamszostak/LLM-Learning-Lab

RAG Overview: Retrieval augmented generation (Stack Overflow)

What is RAG? (Amazon)

Best Practices in Retrieval Augmented Generation (Gradient Flow)

Embeddings: What Are Embeddings (OpenAI)

What Are Word and Sentence Embeddings? (Cohere)

RAG vs. Fine-Tuning:

3 Ways To Improve Your Large Language Model (Maarten Grootendorst)

Fine-Tuning, Prompt Engineering & RAG! (LinkedIn)

hack.diversity