Westies

Citi HackOverFlow 2023

Kang Jun Hui Bryan
Mandfred Leow Hong Jie
Jazlyn Phua Jie Ling
Tan Jing Han
Jeff Tan

OUR TEAM

We are Westies, a team of passionate and forward-thinking individuals who are on a mission to shape the future of technology through innovation, creativity, and a relentless pursuit of excellence.

Jazlyn

Bryan

Han

Jeff

Mandfred

TABLE OF CONTENTS

INTRODUCTION

The problem statement:
There is a need to **centralize and streamline knowledge** management within the banking sector to enhance productivity and decision-making.

PROBLEM STATEMENT

- Scattered Data
- Inefficiencies in Information Retrieval
- Lack of Cohesion
- Need for Instant Answers

TARGET USER GROUP

- General Citibank Customer
- Millenials and Gen Z
- Citibank Staff

SOLUTION

Citibot Charlie

A full-stack chatbot that leverages OPENAI Language Model (LLM) as the core technology behind its natural language understanding and generation capabilities.

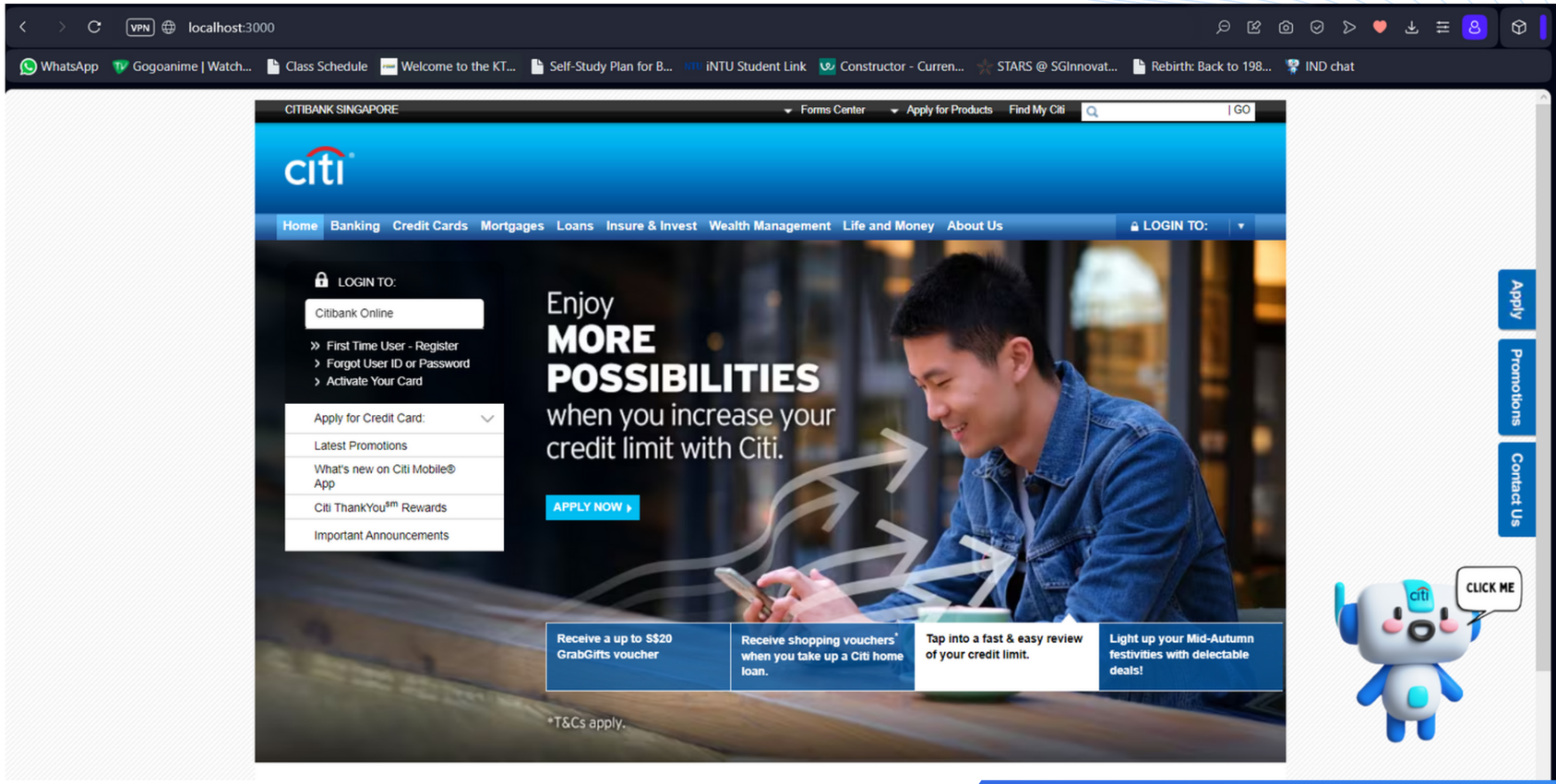
TECHSTACK

REACT	Frontend
NEXT.JS	Frontend/Backend
TYPESCRIPT	Frontend
TAILWIND	Frontend
LANGCHAIN	Python Lib
PINECONE	Database
OPENAI	API(LLM)

KEY FEATURES

CITIBOT CHARLIE

Your friendly website assistant



KEY FEATURES

- Serves as a **one-stop chatbot** for citibank internal usage
- Trained with **customized data** of your choice

MR CHARLIE

Internal Citibank staff assistant

LIVE DEMONSTRATION

TECHNICAL OVERVIEW

- **Pinecone API**

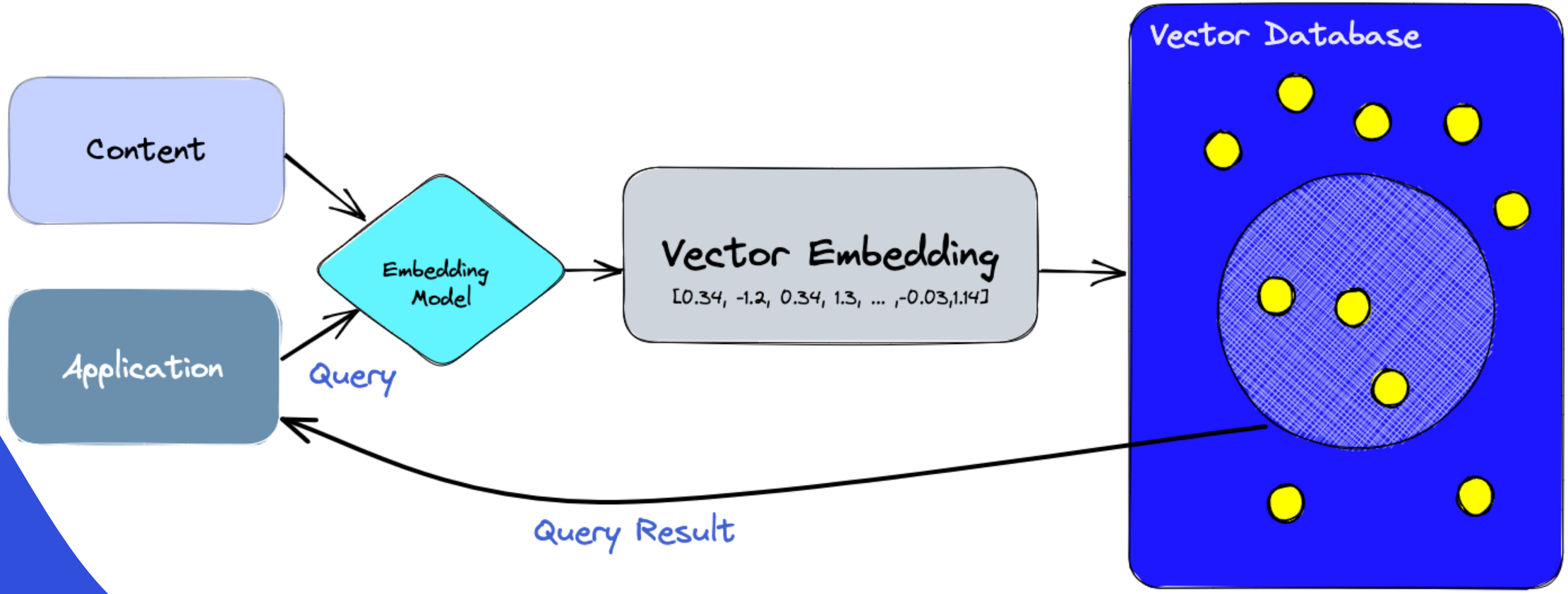


Figure 1. Vector Database. Retrieved from https://www.pinecone.io/learn/vector-database/

TECHNICAL OVERVIEW

OpenAI (LLM)

- **3.5 turbo**

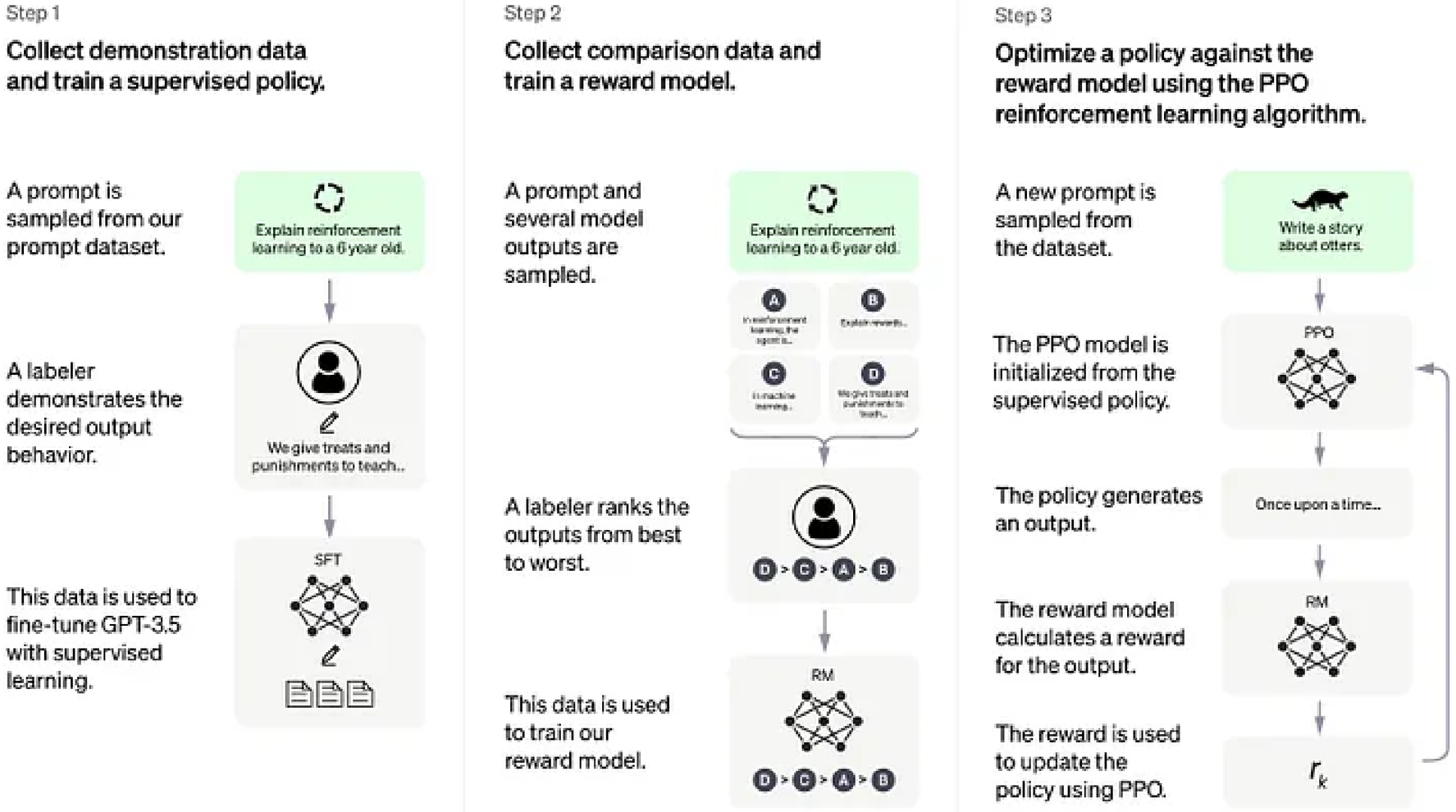


Figure 3. Reinforcement Learning. Retrieved from https://medium.com/@amol-wagh/open-ai-understand-foundational-concepts-of-chatgpt-and-cool-stuff-you-can-explore-a7a77baf0ee3

TECHNICAL OVERVIEW

- **LangChain**

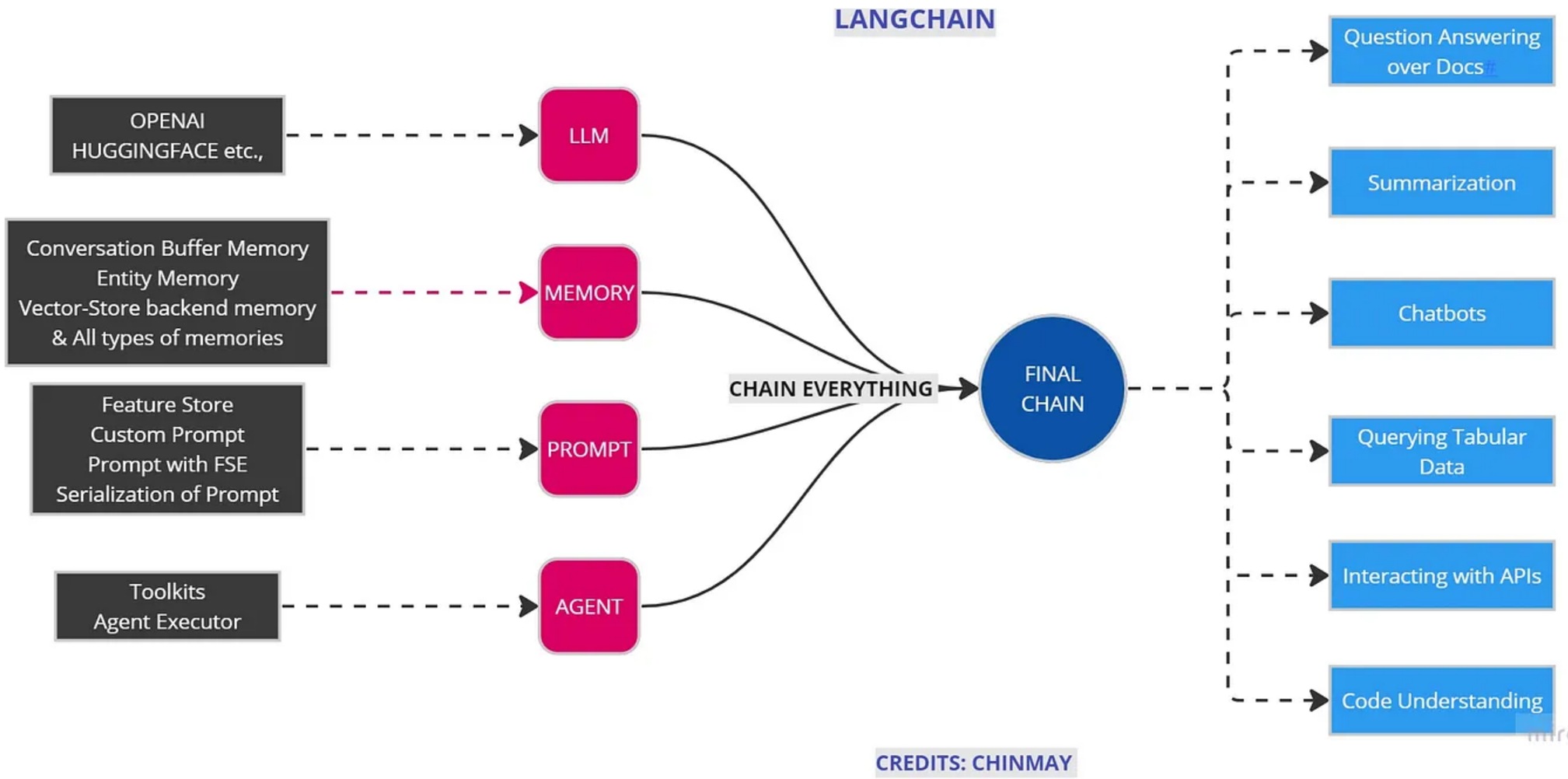


Figure 2. High Level Structure of LangChain. Retrieved from https://pub.towardsai.net/understanding-langchain-%EF%B8%8F-part-1-98499559f4c4

BENEFITS

- **Efficiency**
- **Accuracy**
- **Scalability**
- **Cost-effective**

FUTURE ENCHANCEMENT

INTEGRATION WITH OTHER SYSTEMS

- Core Banking System
- Payment and Transfer Services

NLU IMPROVEMENT

Enhance the chatbot's ability to understand context and nuances in user input.

MULTILINGUAL SUPPORT

Expanding the chatbot to support multiple languages.

MULTIMODAL INTERACTION

Enable voice recognition input and response capabilities to allow multitasking or help those with accessibility needs

CONCLUSION

- Considering the target market, we derive an attractive solution that is efficient, accurate, scalable, and cost-effective.
- Solution is easily integrable to Citibank's main page where there is high website traffic.
- Solution is abstract to be used in most knowledge-management problem

Jazlyn

Y4 NTU IEM

Bryan

Y3 NTU CS

QnA

Han

Y3 NTU CS

Jeff

Y3 NTU CS

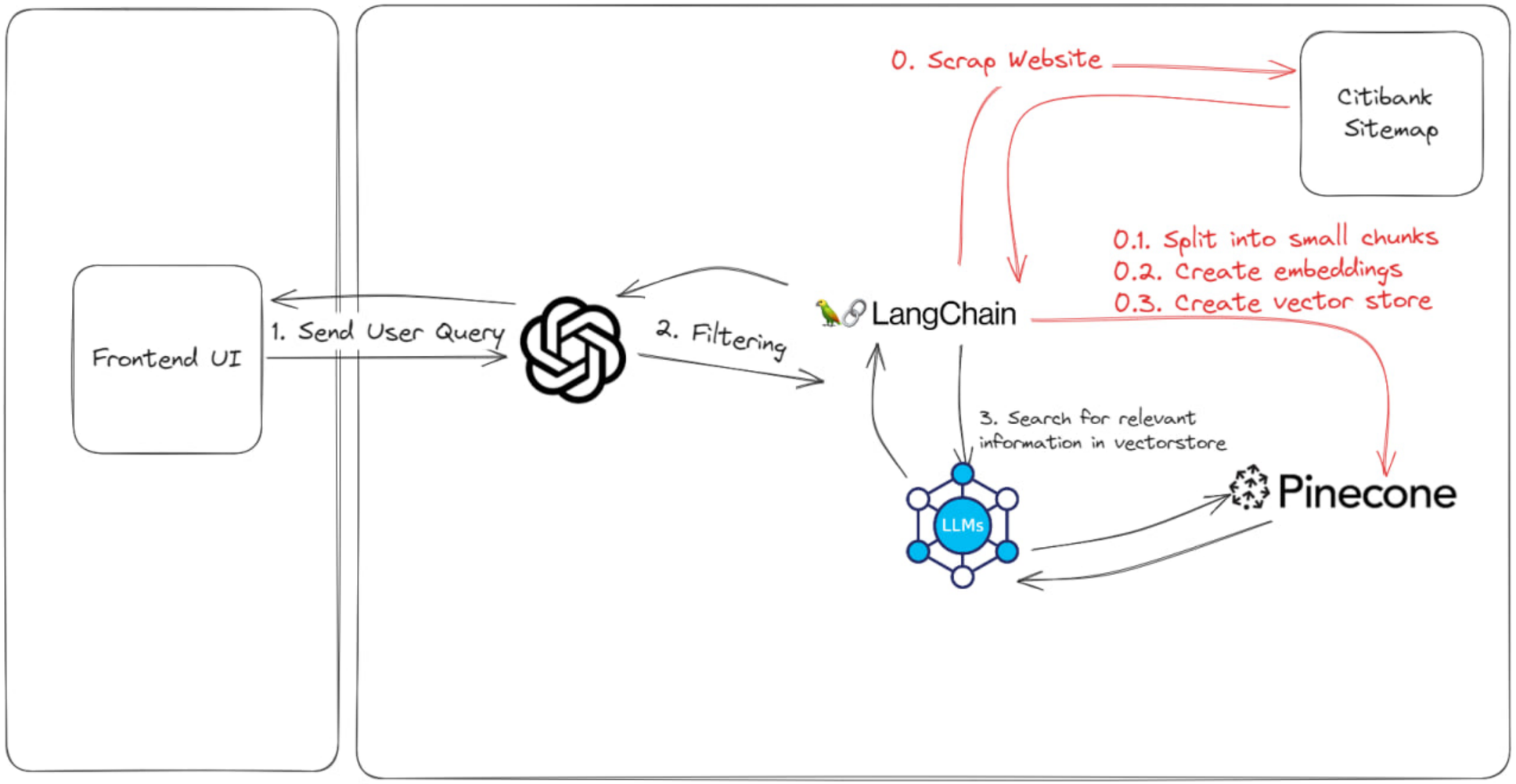
Mandfred

Y3 NTU CS

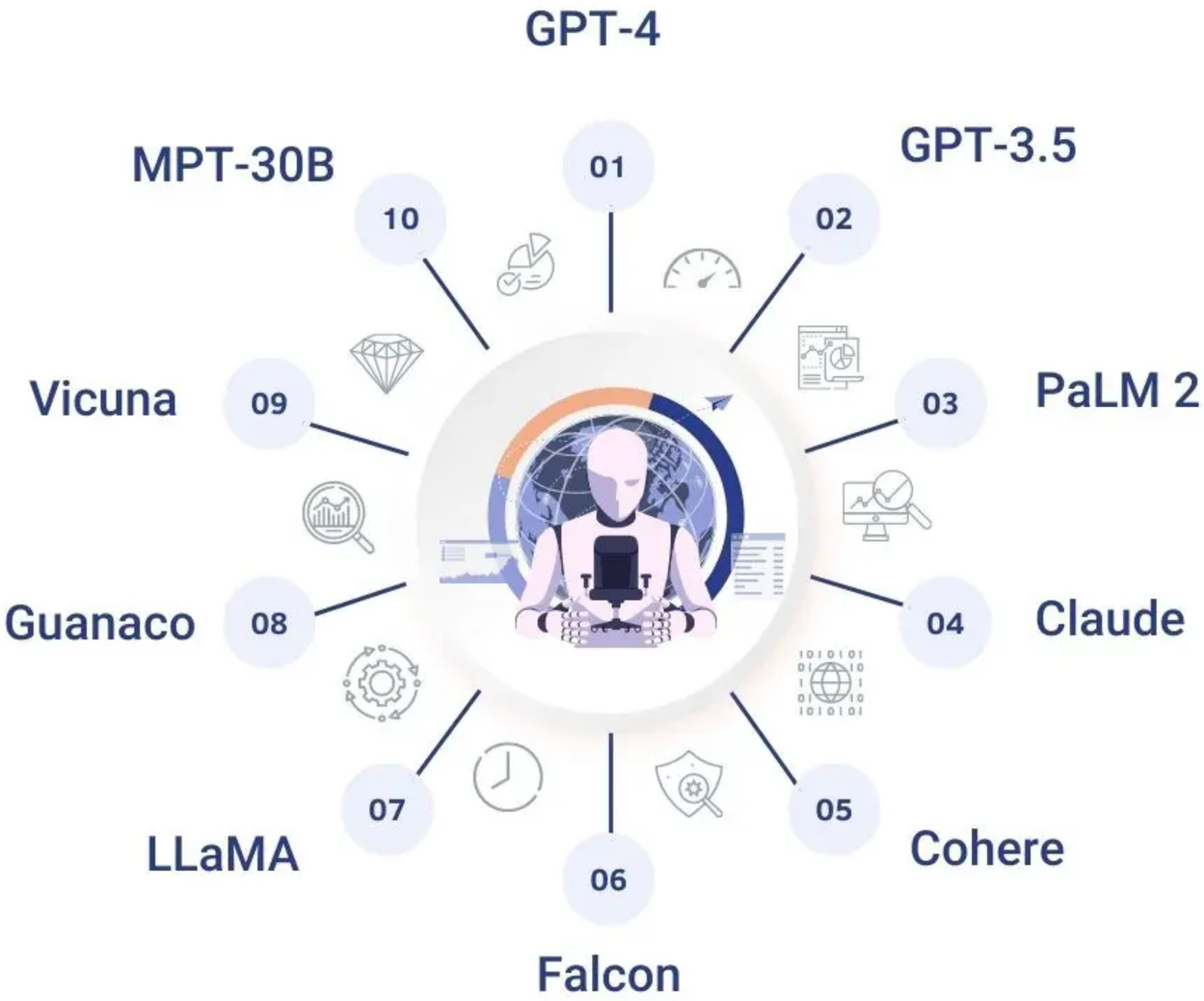
THANK YOU

APPENDIX A : TECH

HOW THE FEATURES WORK?



OTHER TYPES OF LLM MODEL



BENEFITS OF VECTOR STORE?

Semantic Representation: Vectors can capture the semantic meaning of data. For instance, word embeddings can capture the semantic meaning of words or sentences, and image embeddings can capture the content of images. This allows for more meaningful comparisons and operations on the data.

Efficient Similarity Searches: Vector spaces allow for efficient similarity searches, such as finding the most similar items to a given item. This is particularly useful in recommendation systems, image retrieval, and other applications where you want to find items similar to a given query.

Compression: In some cases, representing data as vectors can lead to compression, especially when the original data is large and the vector representation is of a lower dimension.

WHY PINECONE?

Pinecone can swiftly look for related data points in a database by displaying data as vectors so that it may be searched. This makes it ideal for a range of use cases, including semantic search, similarity search for images and audio, recommendation systems, record matching, anomaly detection, and more

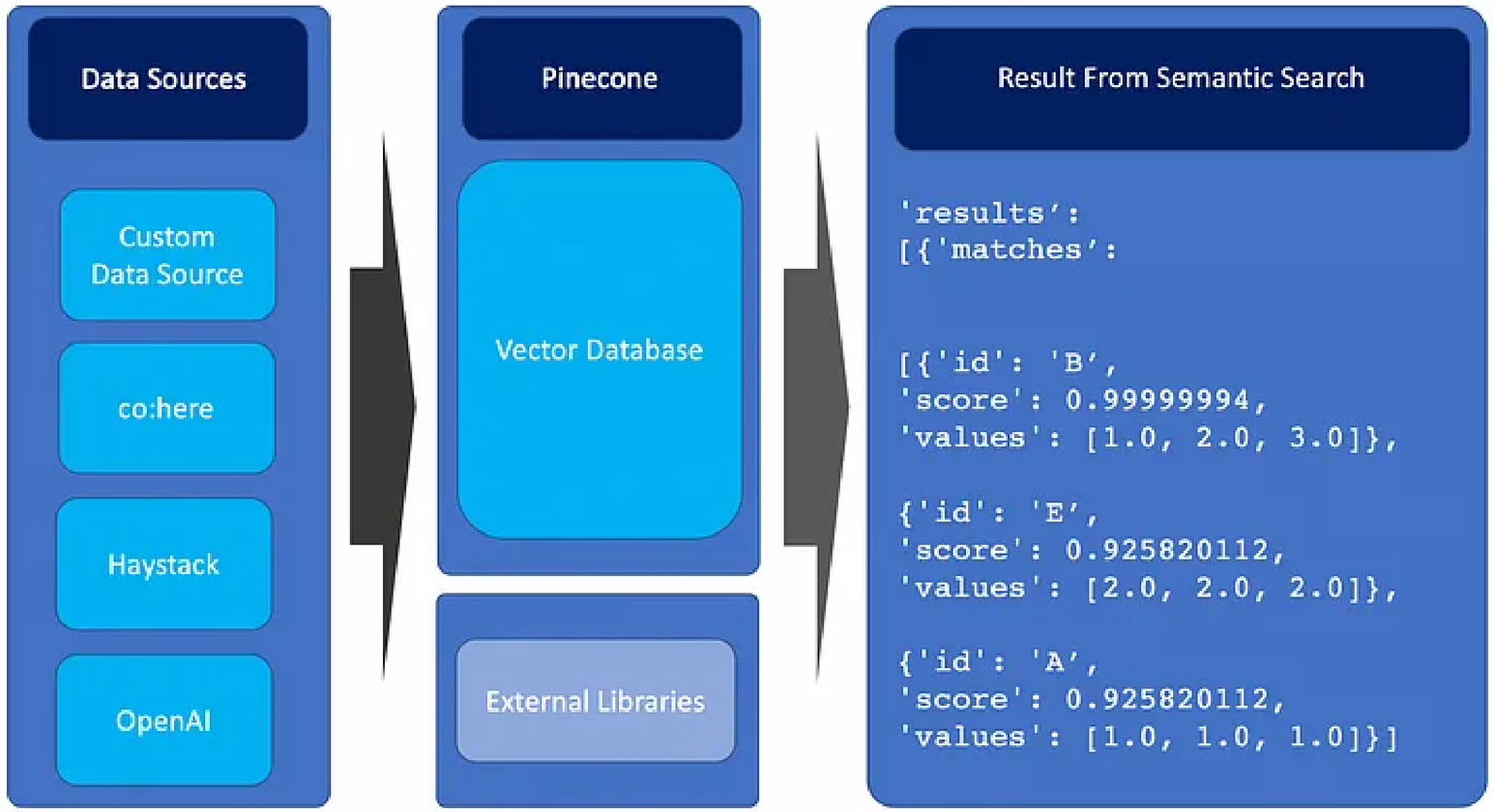


Figure 4. Pinecone process. Retrieved from https://medium.com/@amol-wagh/open-ai-understand-foundational-concepts-of-chatgpt-and-cool-stuff-you-can-explore-a7a77baf0ee3

APPENDIX B : BUSINESS

COST ANALYSIS

OPEN AI	Model	Input	Output
		Usage	
GPT-3.5 Turbo	4K context	$0.0015 / 1K tokens	$0.002 / 1K tokens
	16K context	$0.003 / 1K tokens	$0.004 / 1K tokens
Embeddings models	Ada v2	$0.0001 / 1K tokens	

You can think of tokens as pieces of words, where 1,000 tokens is about 750 words. This paragraph is 35 tokens.

COST ANALYSIS(CONT)

E.g.

1 hour = 750 words

(Inclusive of Input and Output)

24 hours = 18,000 words

(Split them into Input and Output)

OPEN AI	Model	Input	Output	Total
		Usage		
GPT-3.5 Turbo	4K context	$0.0015 * 9 = $0.0135	$0.002 * 9 = $0.018	$0.0315
Embeddings models	Ada v2	$0.0001 * 18 = $0.0018		$0.0018
			Subtotal	$0.0333
Monthly Cost		30 x $0.0333		$0.999

You can think of tokens as pieces of words, where 1,000 tokens is about 750 words. This paragraph is 35 tokens.

COST ANALYSIS (PINECONE)

Pods

An index is made up of pods, which are units of cloud resources (vCPU, RAM, disk) that provide storage and compute for each index. Choose the pod type that works best for your use case.

Indexes

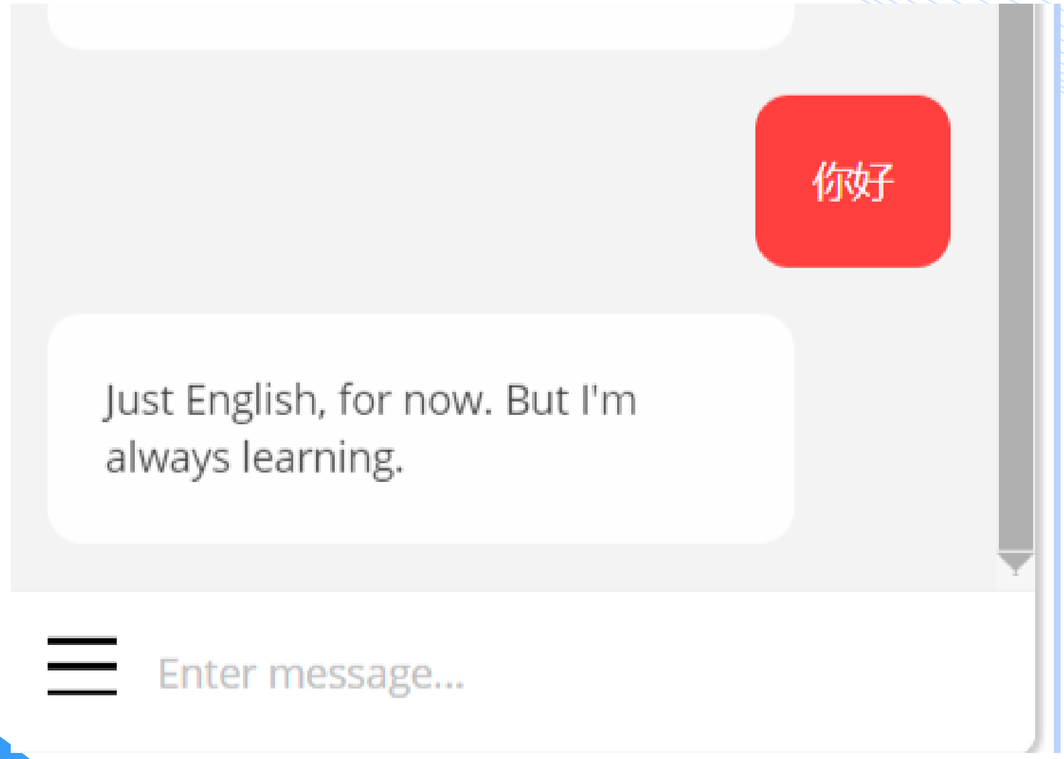
Indexes store your vector embeddings and metadata. Each index uses at least one pod, but you can add more to increase storage capacity.

Scaling

As you grow, you can scale storage capacity by increasing your pod sizes. You can also increase or decrease throughput when needed, by adding replicas to an index.

PLANS	
Standard	**Enterprise**
starting at $70/month	starting at $104/month
For production applications at any scale.	For mission-critical production applications.
Standard Support	Premium Support

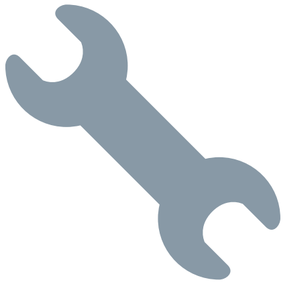
COMPETITOR ANALYSIS



Some chatbot on some website

APPENDIX C : DATA

HOW DID WE GET THE DATA? (CITIBOT CHARLIE)



We have a tool that fetches website data for us.
Think of it like a robot that visits a website, looks at its map (sitemap), and then collects information from all the pages listed on that map.
This robot can collect data from multiple pages at once, but it's polite and doesn't overwhelm the website. It collects data from 2 pages every second.

Setting Up the Environment

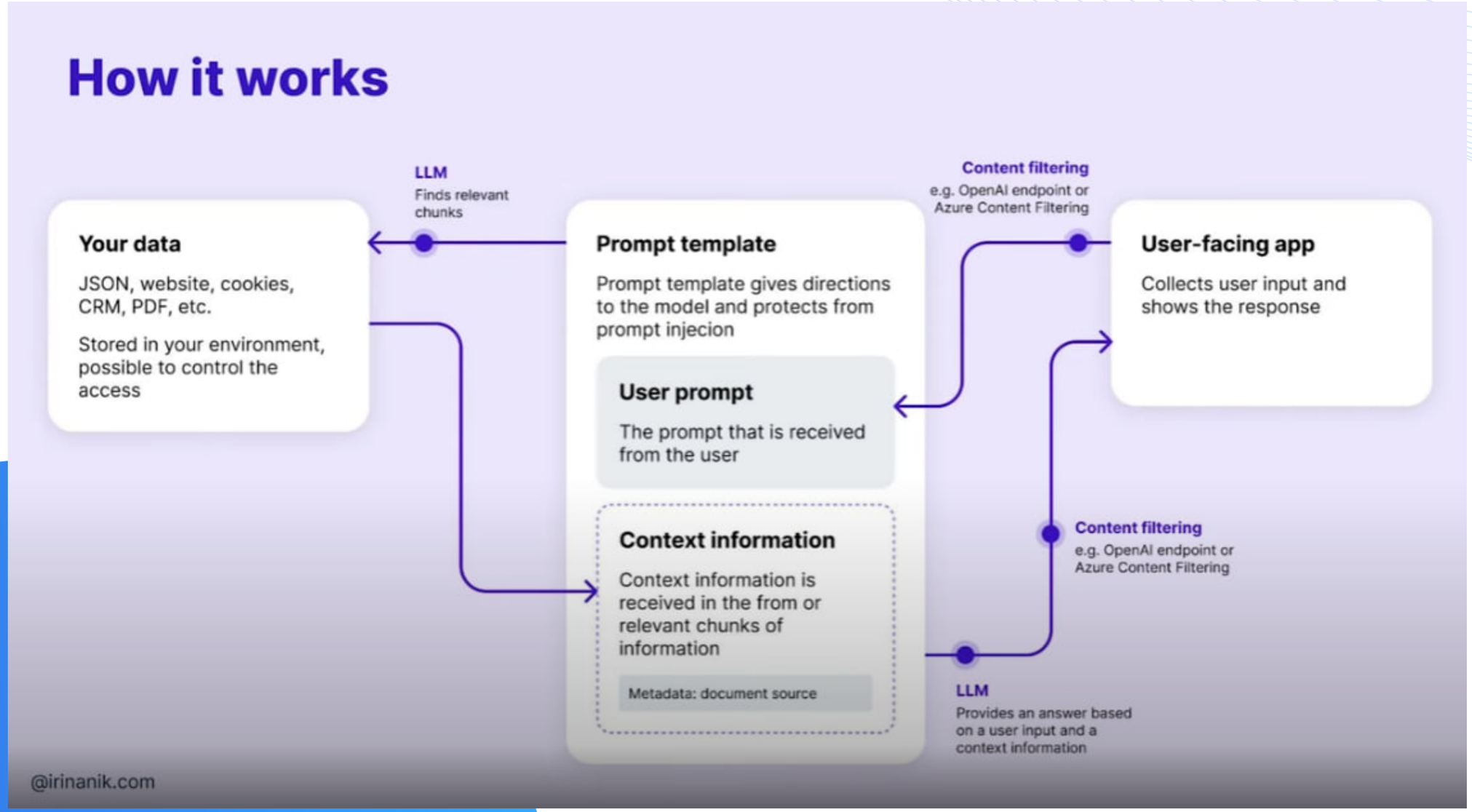
Once we have the data, it's often too big to work with. So, we break it down into smaller, manageable pieces.
Imagine having a long book and dividing it into chapters.

Understanding the Data

Next, we use a tool to understand and represent our data in a way that machines can easily process.
This is called creating "embeddings".
Think of it as translating a language you don't understand into one you do.
Storing and Searching the Data

Now, we need a place to store this translated data and a way to search through it quickly.
We use a tool called Pinecone to do this. It's like a library where each piece of data is a book. We can quickly find the book we want based on its content.
If we're using Pinecone for the first time, we create a new shelf (index). If we've used it before, we just add our new books to the existing shelf.

HOW IS DATA TRAINED?



TYPE OF DATA COLLECTED

- Json Format for the vector
- Data can be of any nature
  - text,images,audio,video,Tabular Data