VNUHCM - University of Science

INTRODUCTION TO BIGDATA

ML7 - A TRIPLE OF LLMS: GPT-3, MEGATRON-TURING NLG, AND GEMINI

PRESENTATION - 2024

By group: KHMT2-05

About us

Team member

Trần Hải Phát	21127131
Võ Thanh Hiệp	21127275
Hoàng Nhật Quang	21127543
Hồ Thế Phúc	21127670

Learn from our topic

Get a comprehensive view of LLM models and their applications

Table of contents

Overview

Understand about LLM, its architecture and provide solutions

GPT-3

GPT stands for "Generative Pre-trained Transformer," which is a widely used deep neural network architecture

Megatron-Turing NLG

Megatron-Turing NLG is a powerful natural language generation model developed by NVIDIA

Gemini Pro

Gemini Pro is a powerful AI model from Google, designed for a wide range of tasks across text, code, and other formats

Our application description

Using Streamlit and Pandasai, we built a web-app to demonstrate the implementation of LLMs for the proposed solutions

Demo

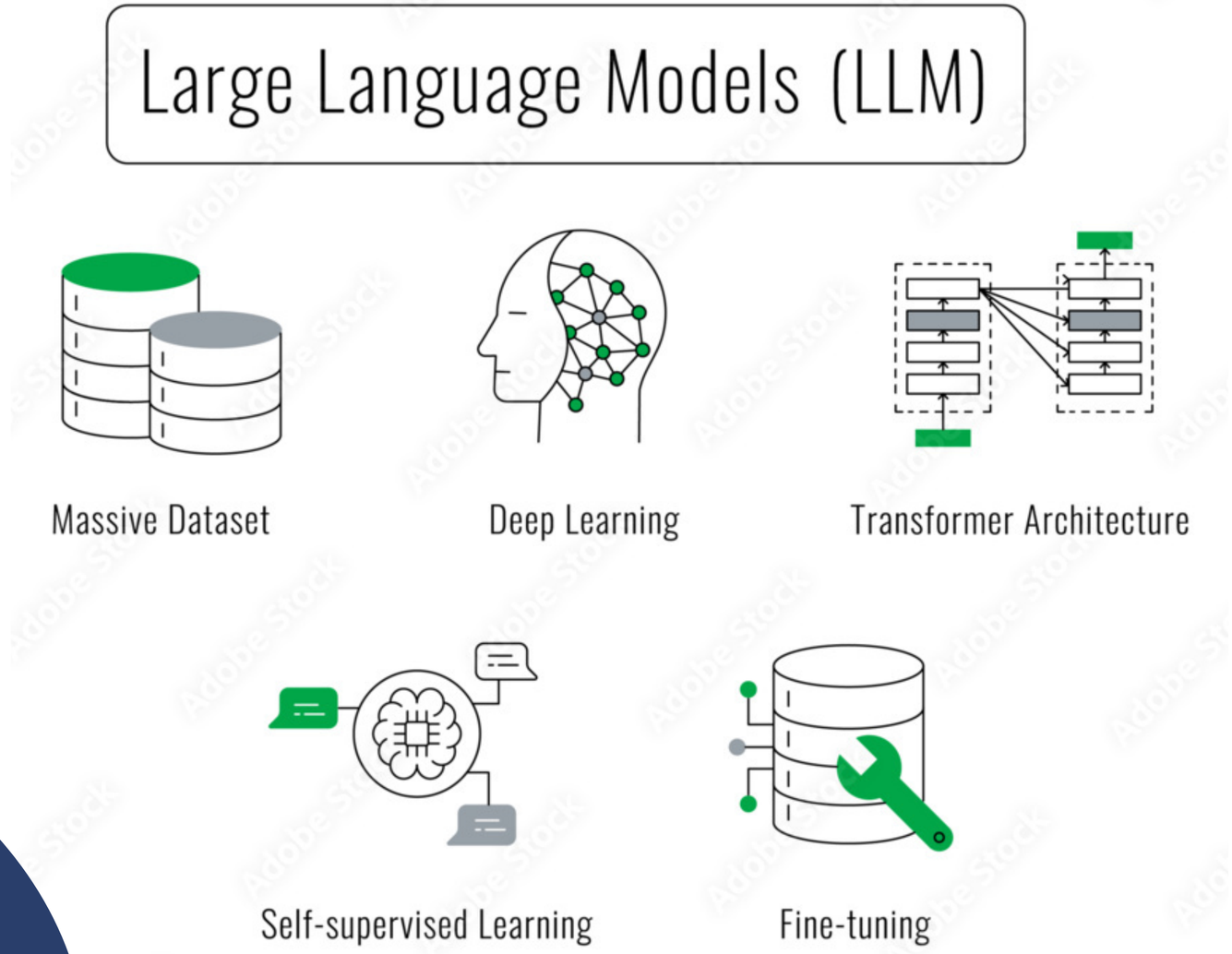
Data chat, data analysis and data visualization

01

Overview

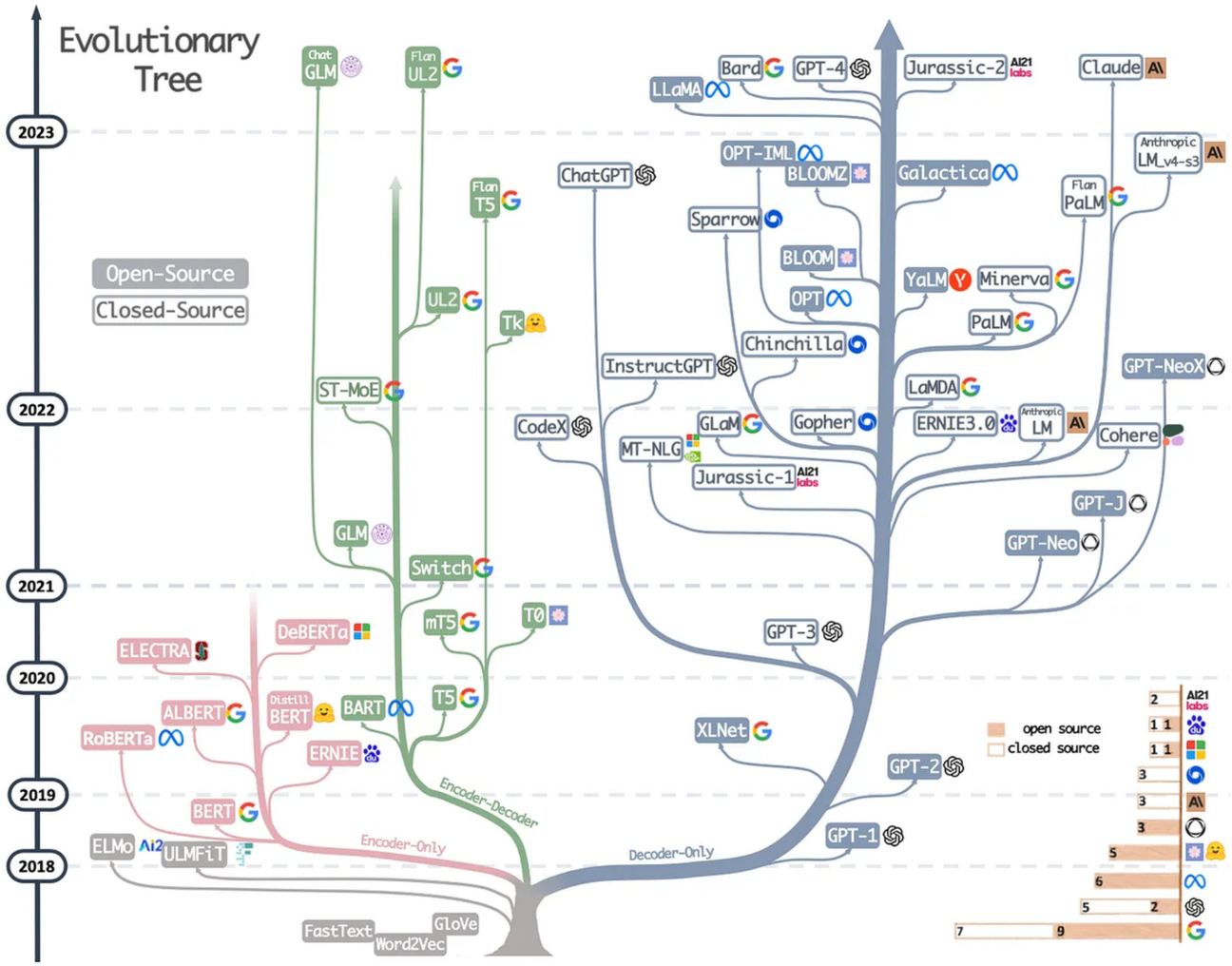
OVERVIEW - Base Knowdlege

What is a large language model (LLM)?



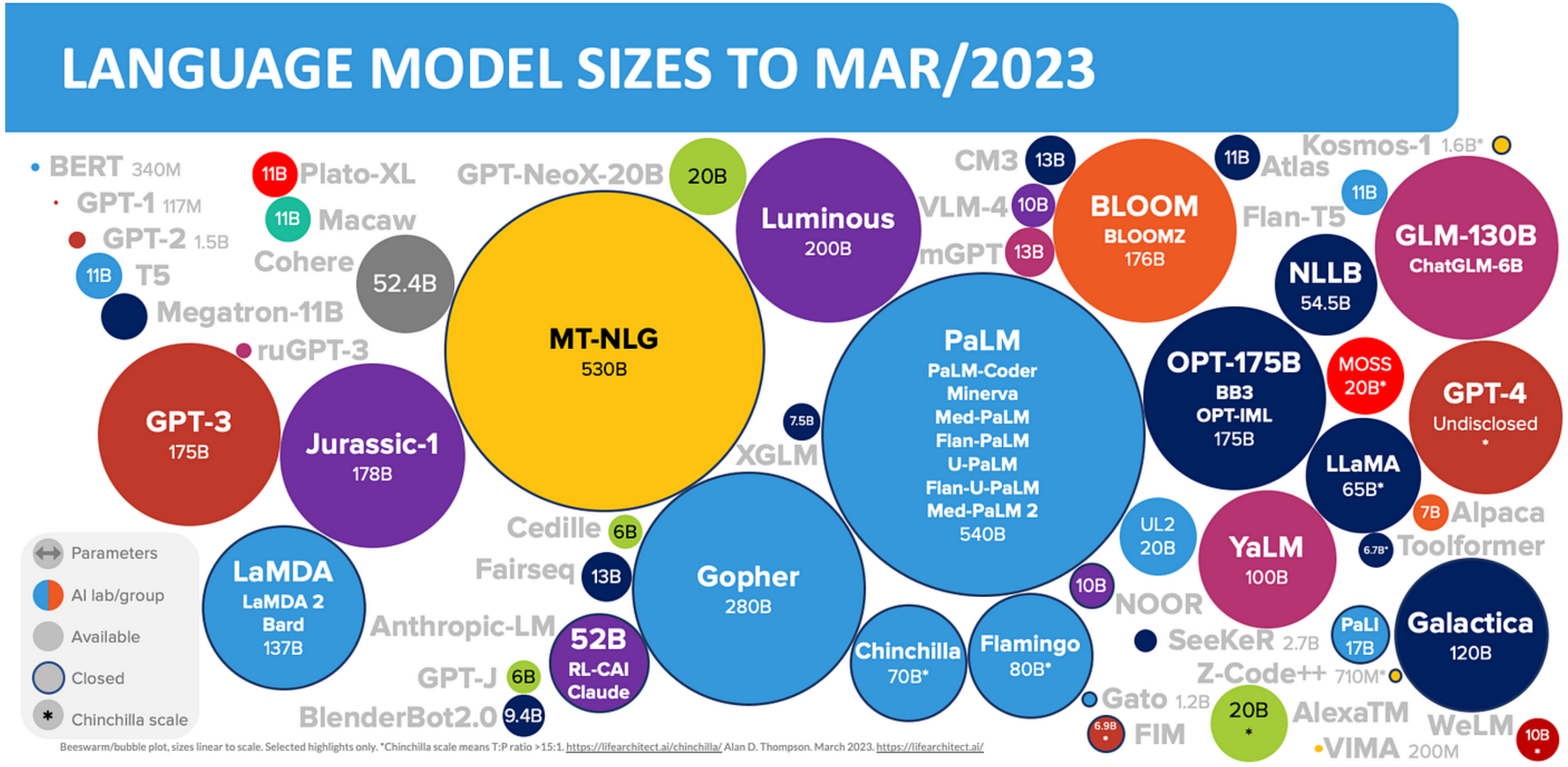
OVERVIEW - Base Knowdlege

The development of language models in recent years



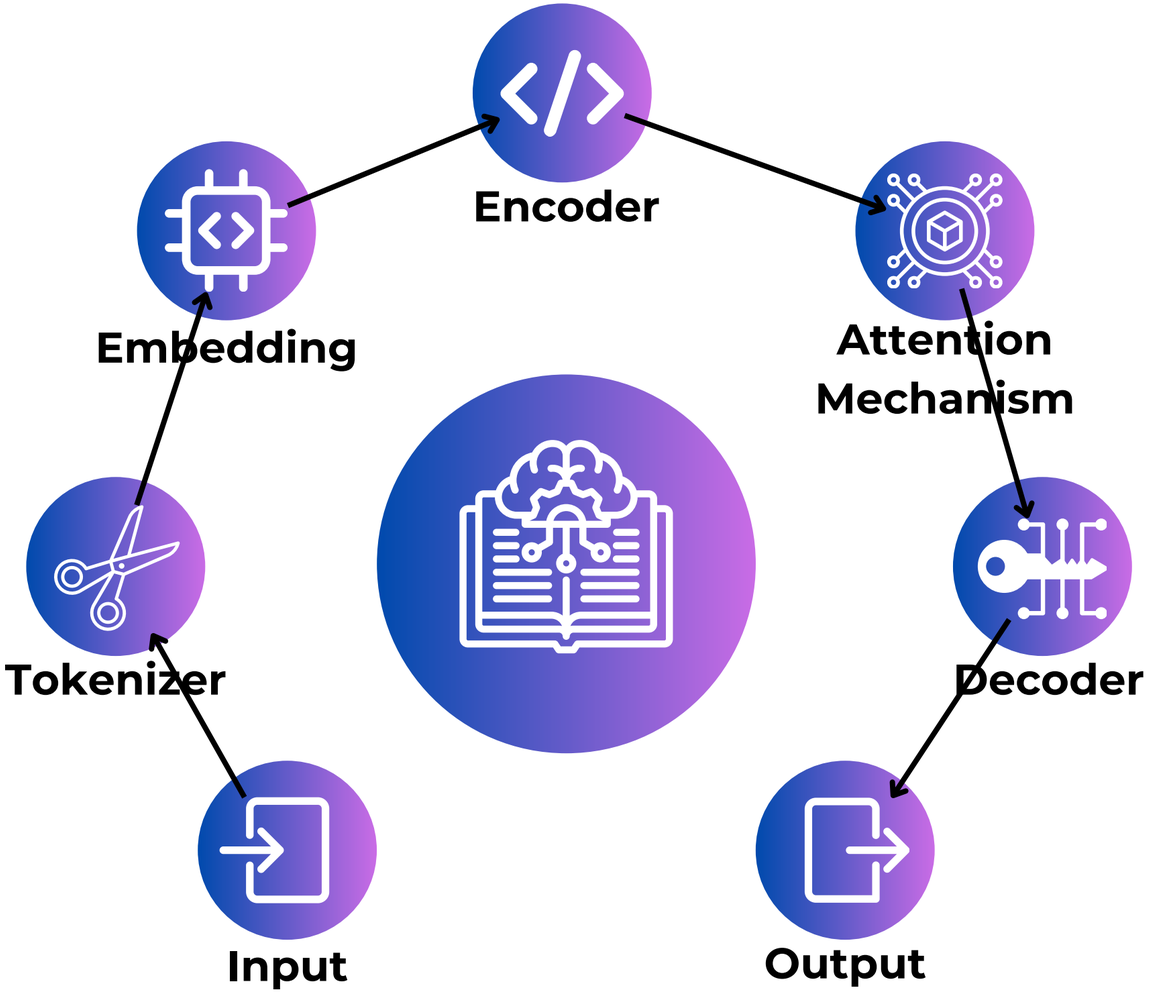
OVERVIEW - Base Knowdlege

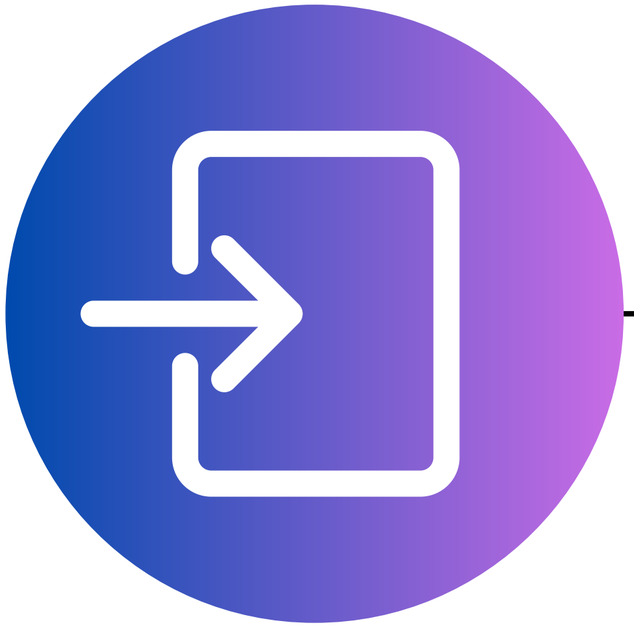
The current Arms Race of LLMs



OVERVIEW - Base Knowdlege

How does LLM work?





Input

Hello, I am a sicentist .

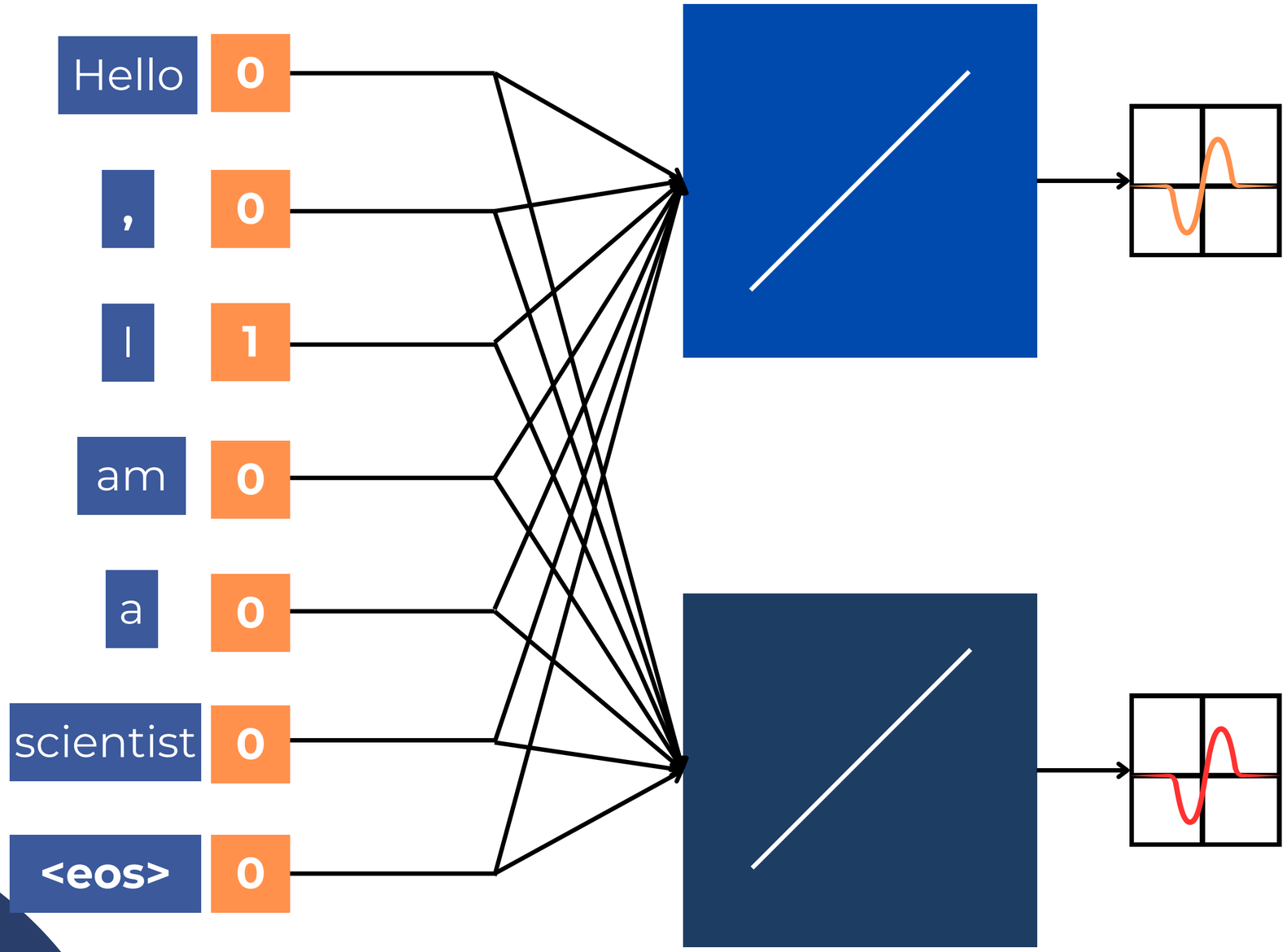
Tokenization

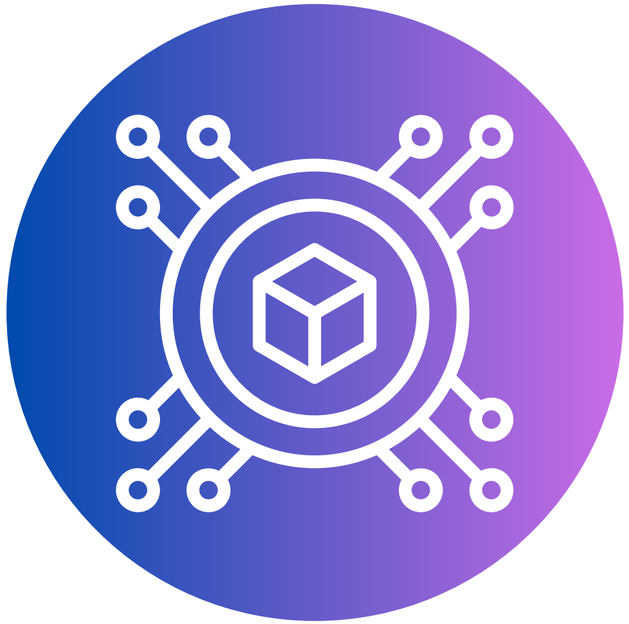
Hello | , | I | am | a | scientist | **<eos>**

Encoding

and

Embedding


Hello
0
,
0
I
1
am
0
a
0
scientist
0
<eos>
0



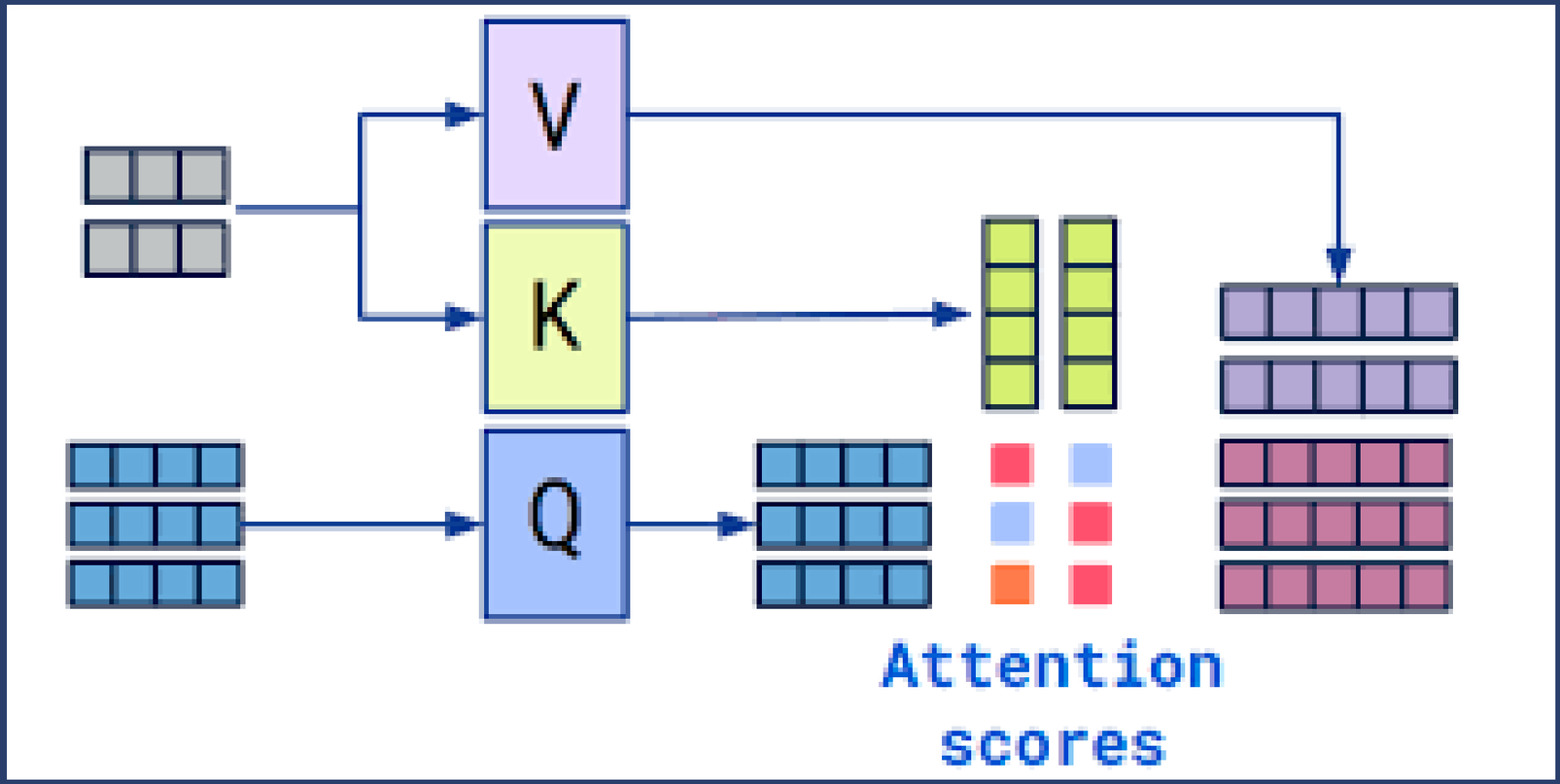
Self-attention mechanism

Masked Self-attention

Example sentence:
I go to HCMUS in District 5, It have Chinese food.

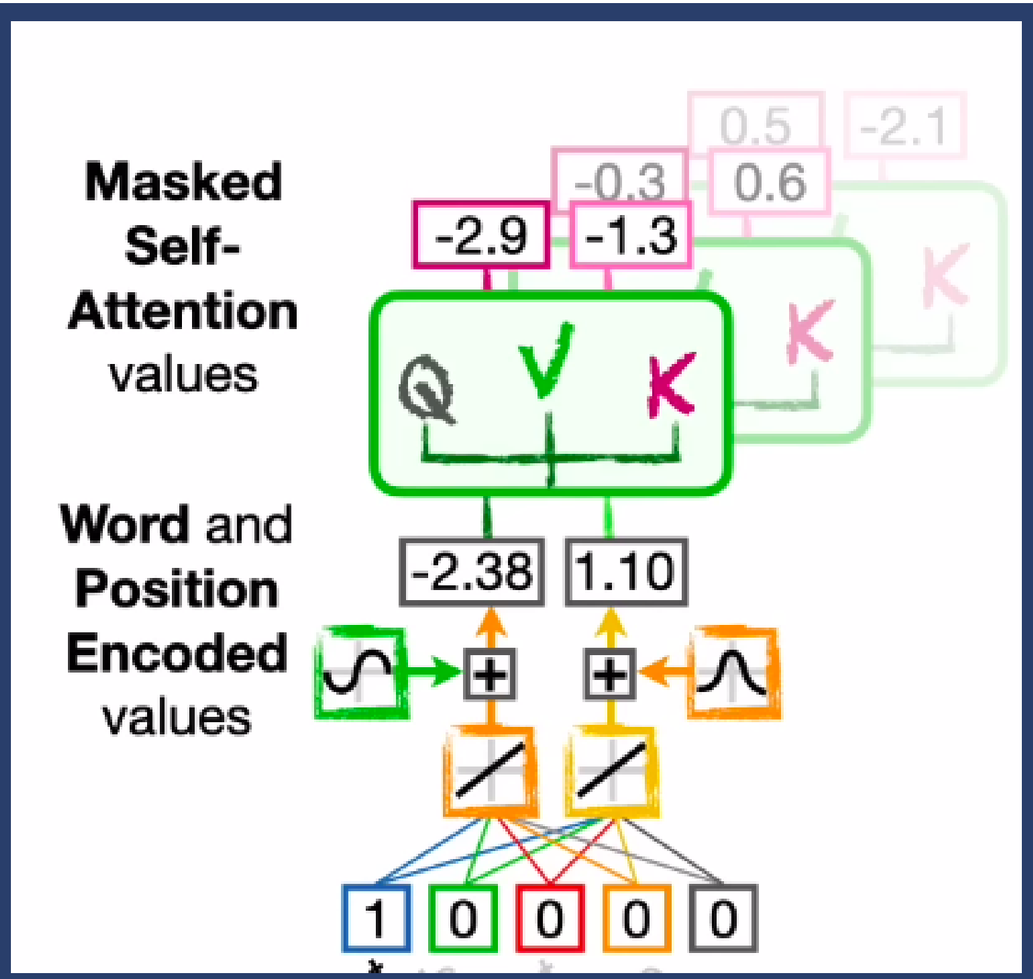
Masked self-attention calculate similarity between a word and all preceding word before it

Masked Self-attention

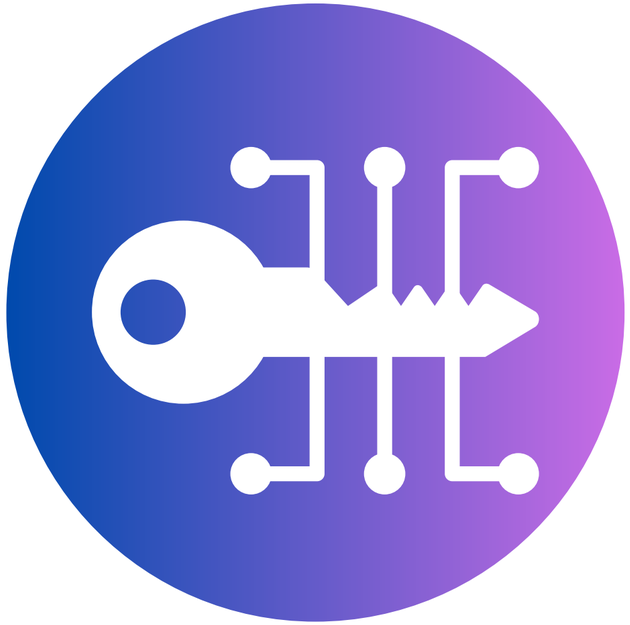


Multi-head Self-attention

key



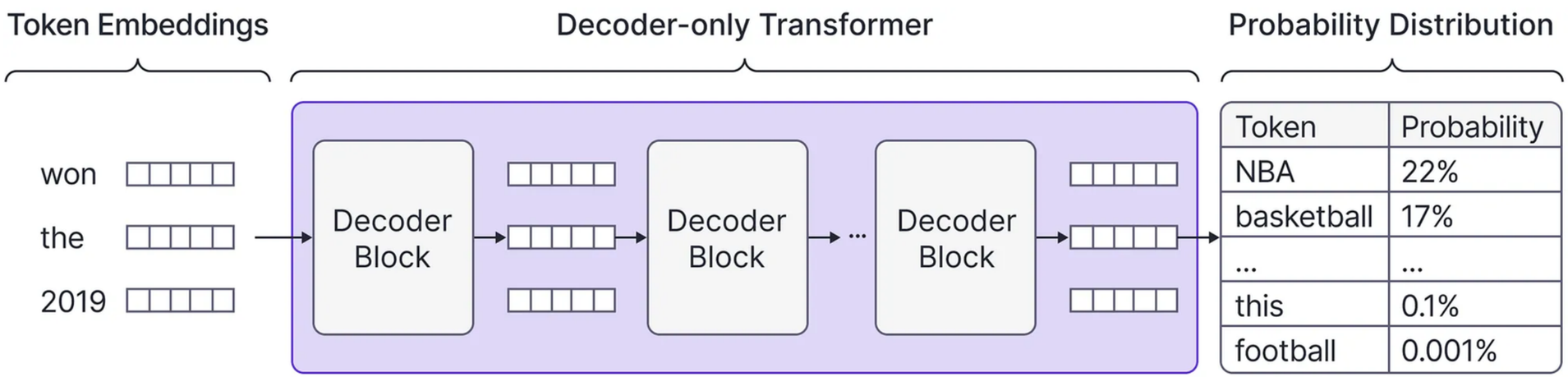
query

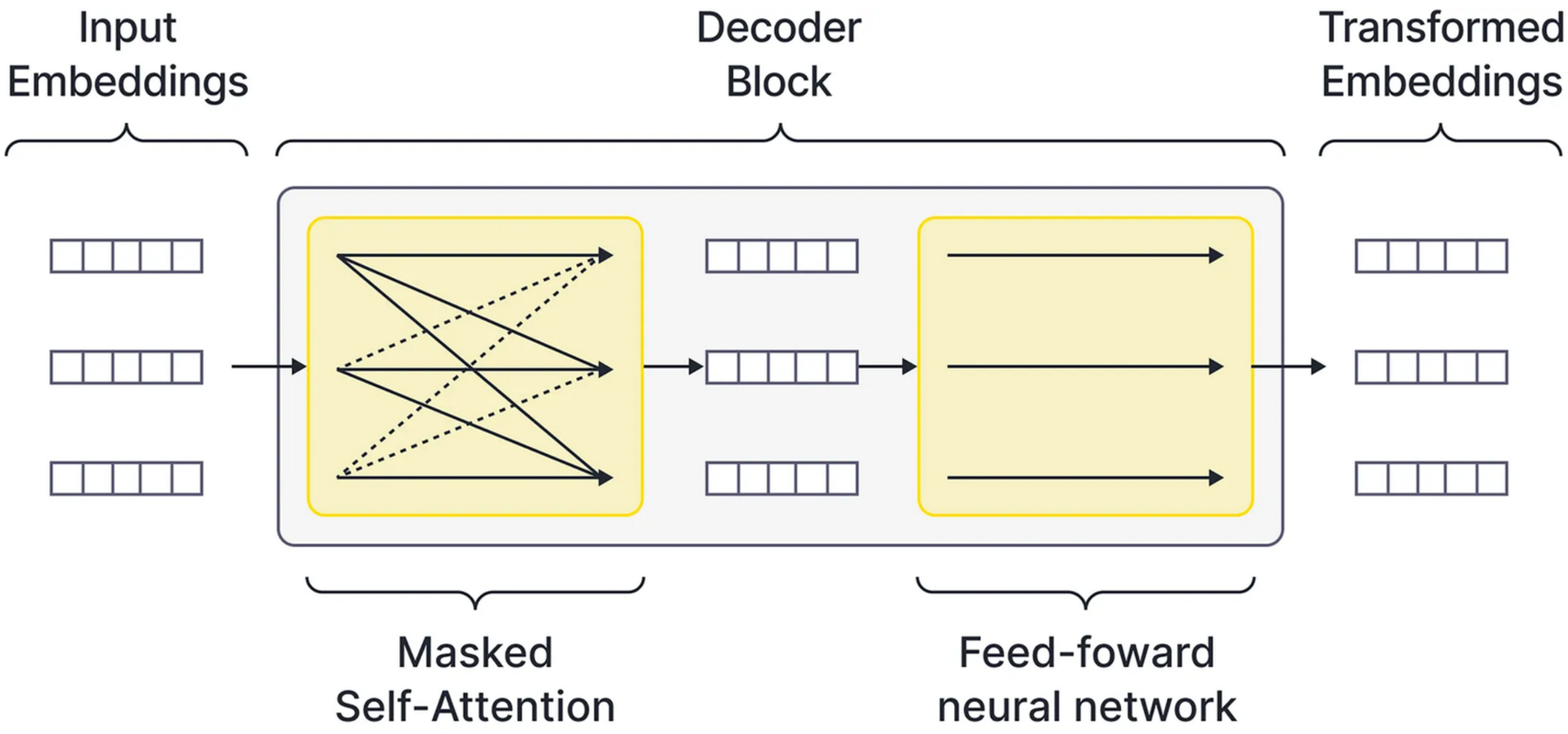


Decoder

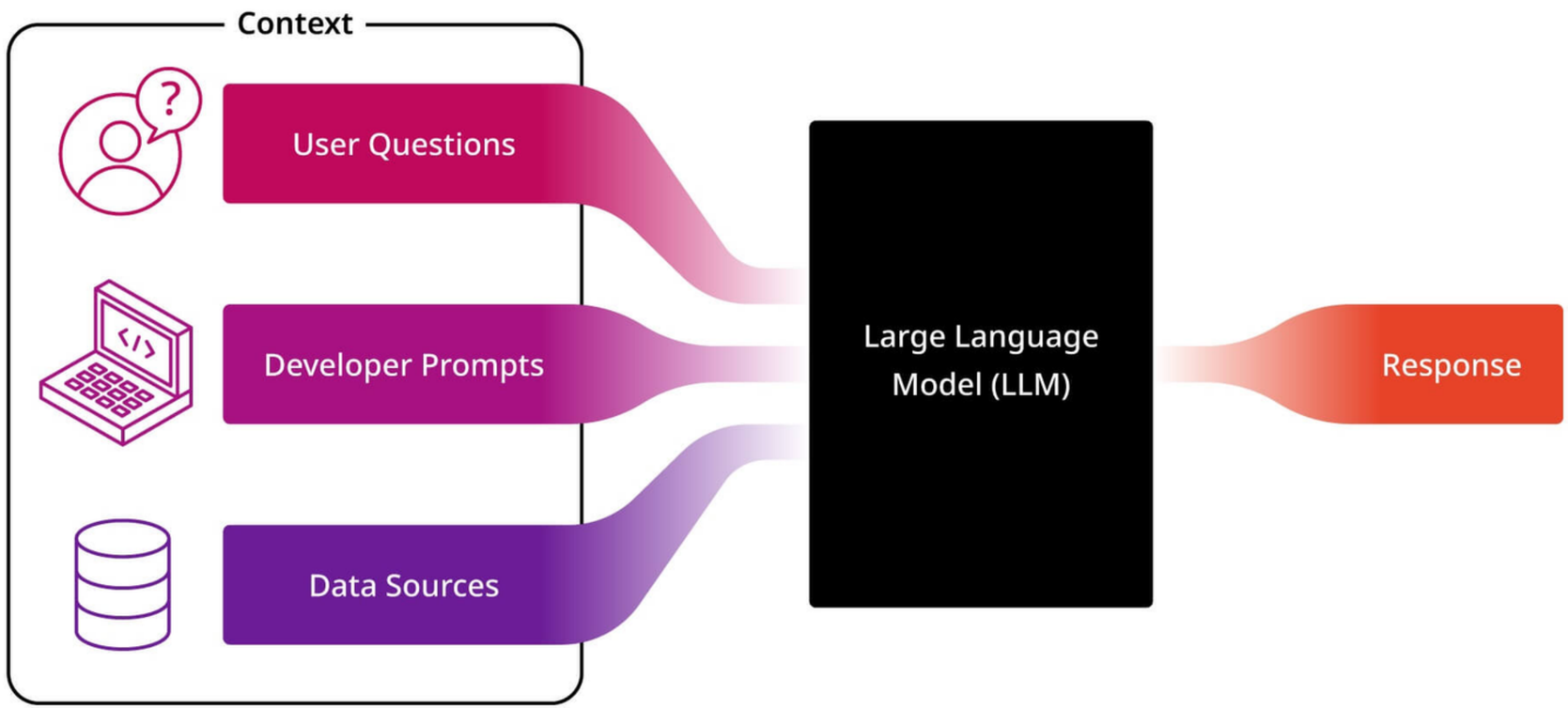
and

Output


Token Embeddings
Decoder-only Transformer
Probability Distribution
won
the
2019
Decoder Block
Decoder Block
...
Decoder Block
Token
Probability
NBA
22%
basketball
17%
...
...
this
0.1%
football
0.001%


Input
Embeddings
Decoder
Block
Transformed
Embeddings
Masked
Self-Attention
Feed-foward
neural network

Overview - Solutions



Overview - Solutions

Usage: Utilize large language models (LLMs) in building a robust question-answering system to perform data chat, data analysis and data visualization.

Data chat

Process natural language queries related to data, extract relevant information from datasets, and provide clear answers and insights to users

Data analysis

Use for text analysis tasks like sentiment analysis, anomaly detection, and identifying unusual patterns in financial news, social media mentions, and other textual data sources

Data visualization

Combine the insights from the LLMs with visualizations to create interactive dashboards for data-driven decision-making

Overview - Solutions

Purpose

- Improve efficiency, accuracy, and customization in processing and deriving insights from large datasets.
- Enable efficient extraction of insights from Big Data sources through robust question-answering systems powered by LLMs.
- Implement scalable information retrieval methods to summarize and provide quick access to relevant information from voluminous data repositories using LLM technology.

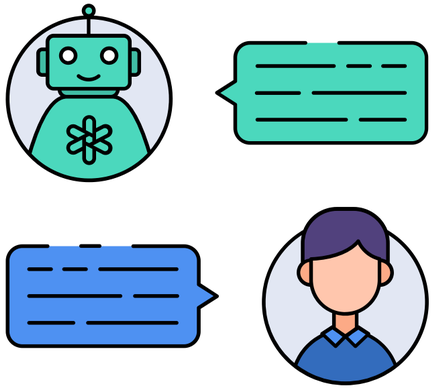
02

Topic Description

GPT-3

Model Size

175 billion parameters



Performance

Strong performance on a variety of NLP tasks without fine-tuning

Training data

Trained on a diverse range of internet text such as common crawl, webtext2, Books1, Books2, Wikipedia

Applications

For tasks like content generation, enhancing chatbots, language translation, educational support, healthcare documentation, programming assistance, creative arts, research analysis, and marketing.

Megatron-Turing NLG

Model Size

530 billion parameters

Performance

Is adverties to have overall better performance compared to original goal.

Training data

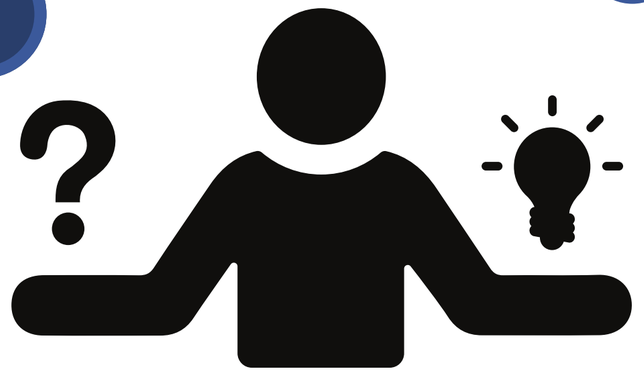
Trained on a diverse dataset called 'The Pile'
stored at least 886GB.

800+

Applications

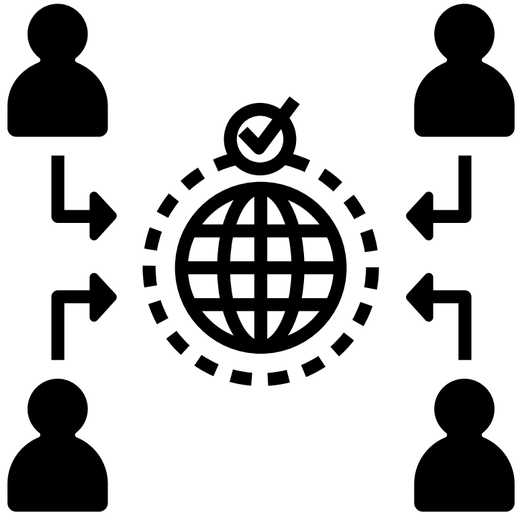
Similar to GPT-3, MT-NLG could be used to generate different creative text formats like poems, scripts, musical pieces, or even code.
However, its potential for more complex reasoning might allow it to create even more intricate and nuanced creative content.

Gemini AI



Model Size

Around 100 billion parameter



Performance

Strong in resoning, understanding and comprehending tasks

Training data

Using reinforced training with pre-train from the start with different modalities. Then fine-tuned with additional multimodal data

Applications

Gemini AI can generate respond like GPT does. But with information being updated continuously, Gemini can provide latest updated responces.
Easiest to get acces to in all **3**.
More humane-like respond.
Generating image.

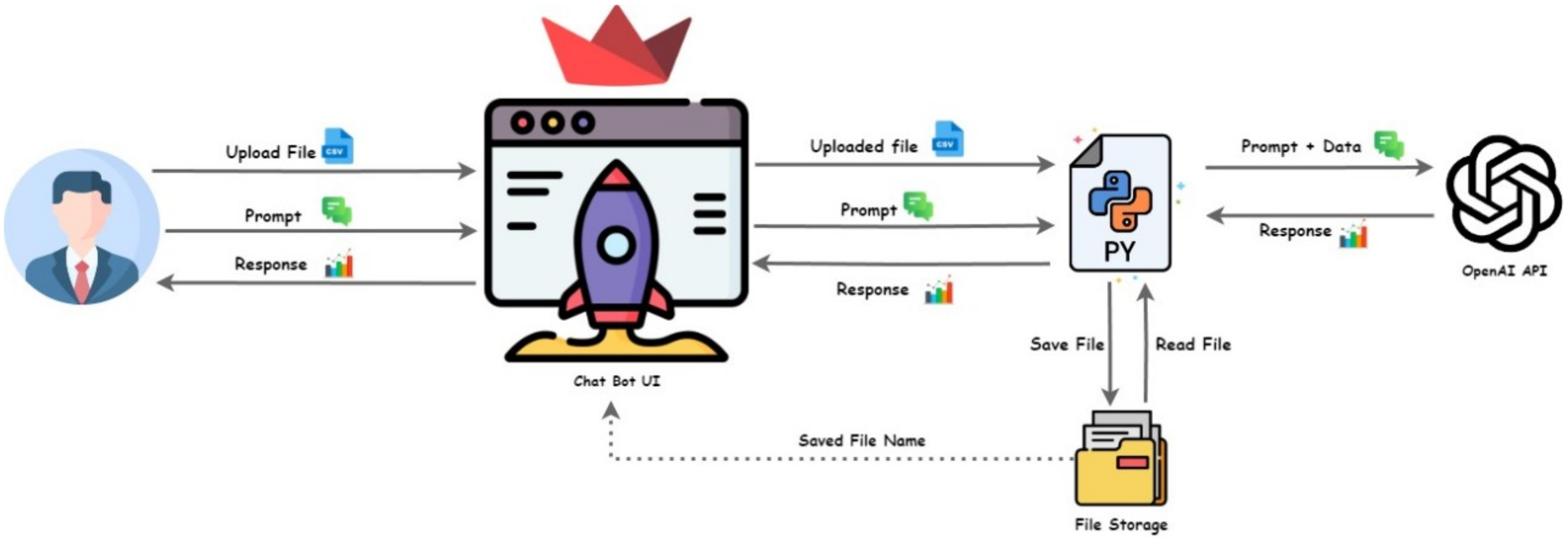
Other LLMs

MODEL NAME	BERT	LLama
MODEL SIZE	BERTBASE is 110 million parameters BERTLARGE is 340 million parameters	65 billion parameters
PERFORMANCE	Currently no clear statement on clarified paper	With 13 billion parameters, it can outperform chatGPT-3
TRAINING DATA	A pre-trained on a massive dataset of unlabeled text	A mixture of several sources in which are also been used in other LLMs.
ACCESSIBILITY	Currently no way to access it except for importing library Transformers from Hugging Face.	There is an open source provided for META in which we have to build the model by ourself

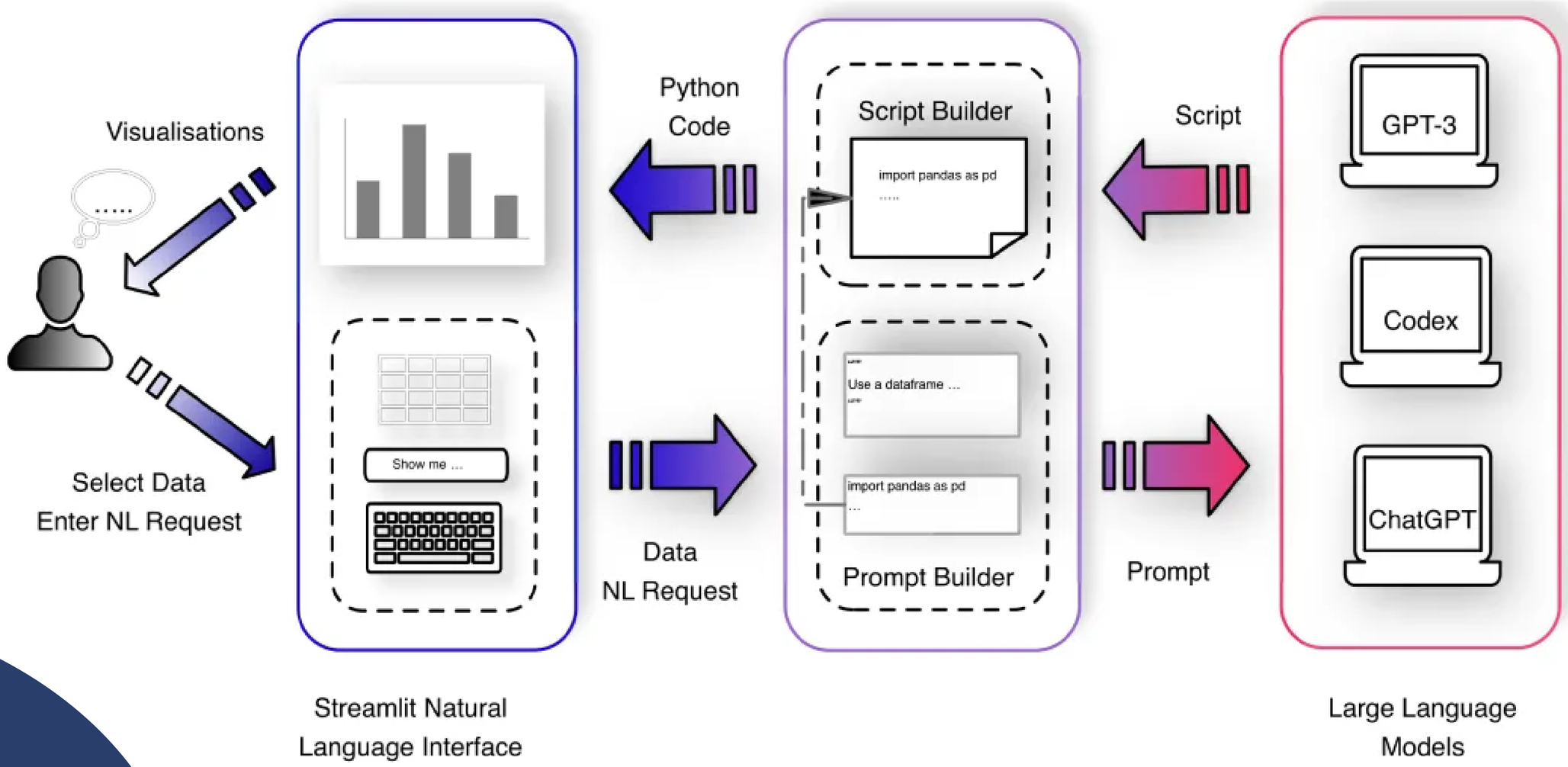
03

Web-app Application

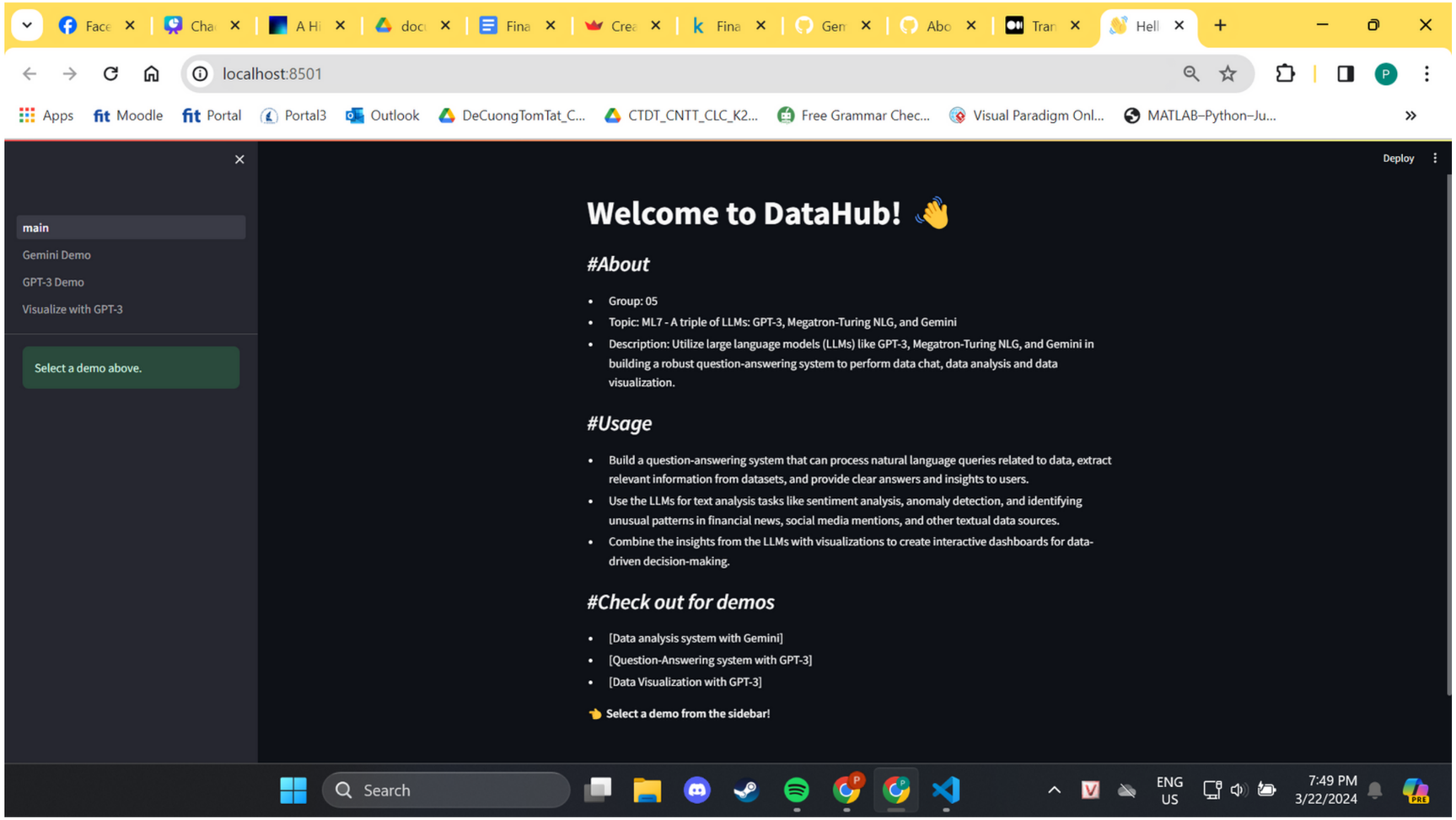
Data Chat & Analysis workflow



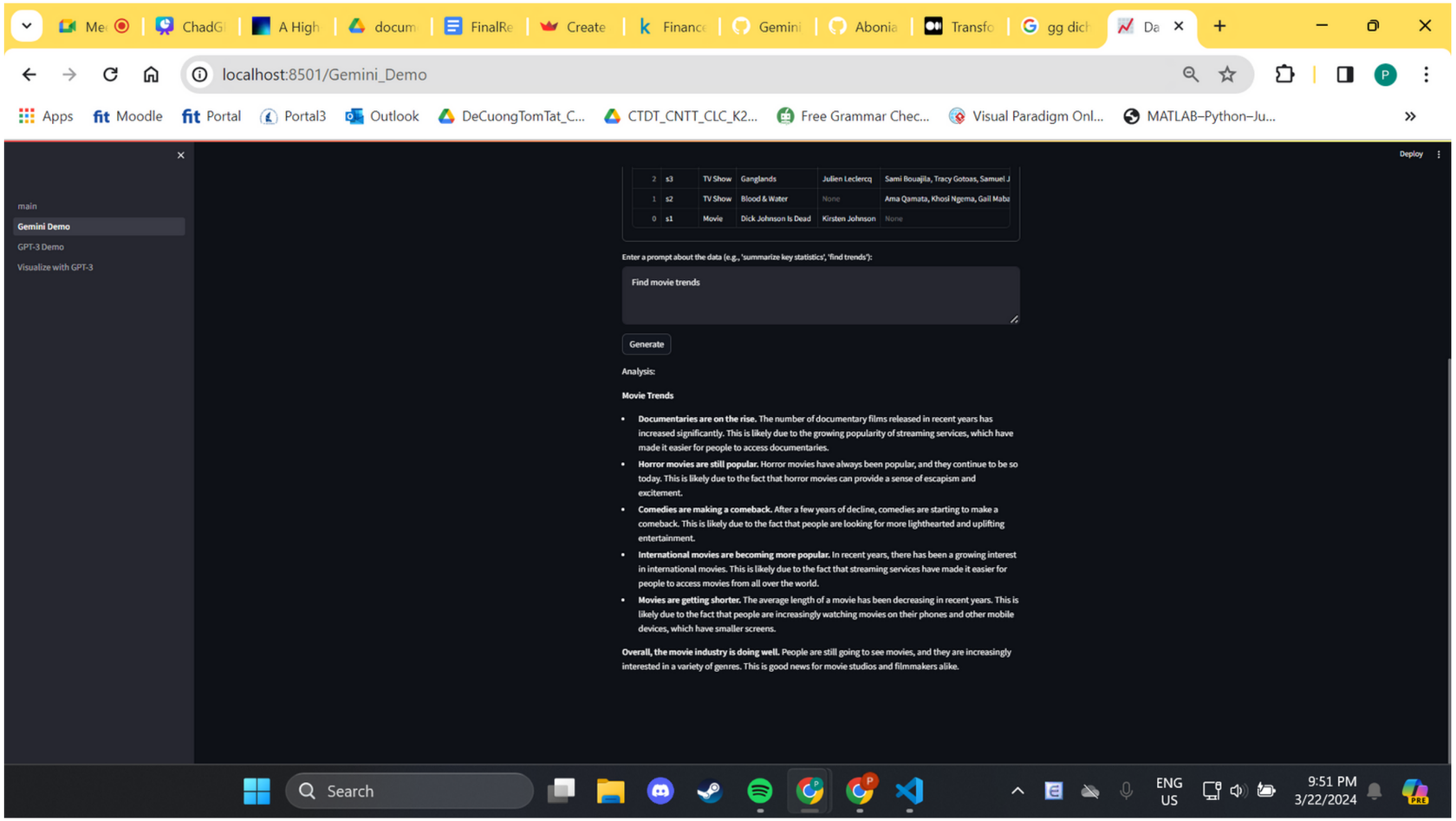
Data Visualization workflow



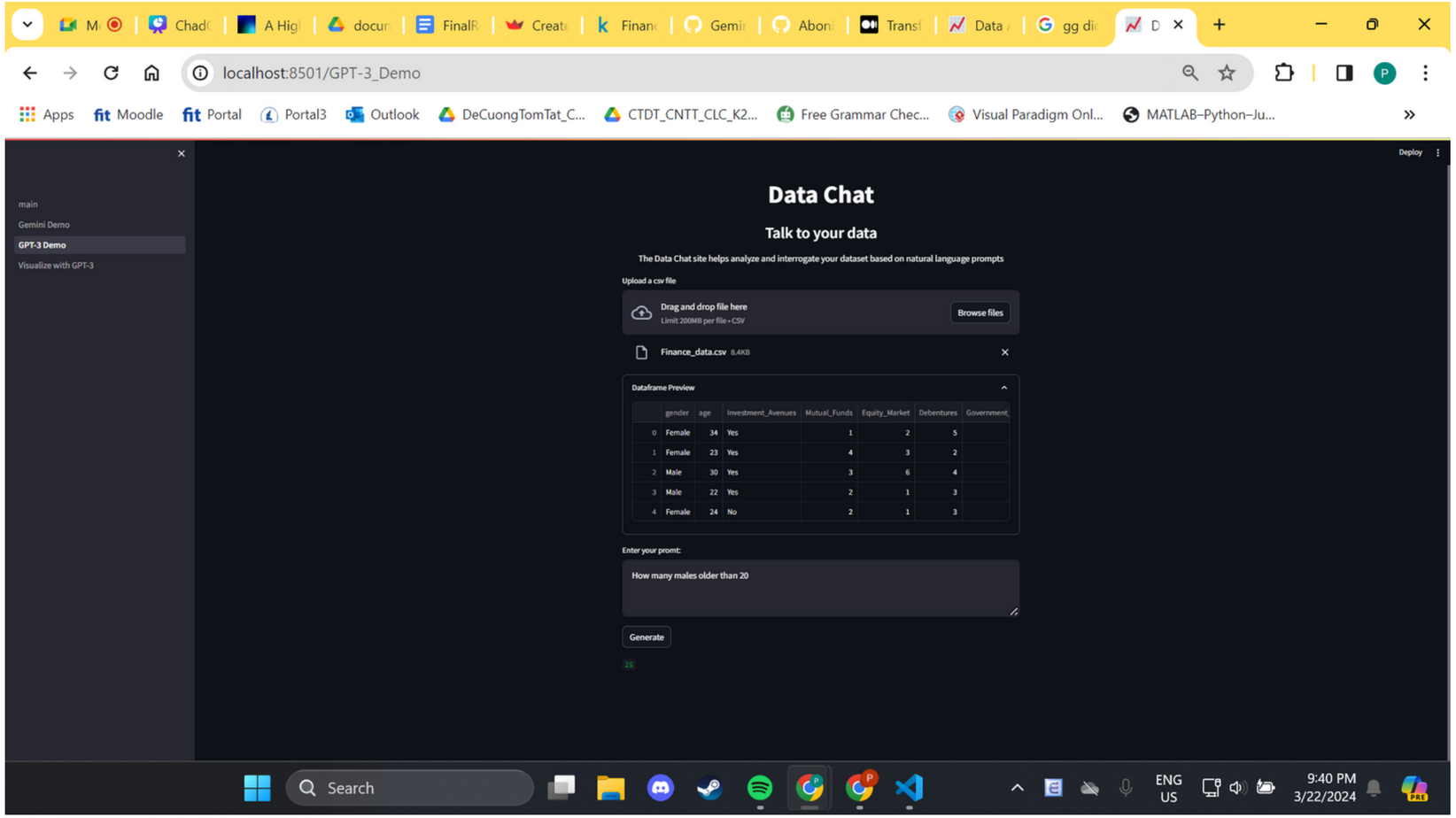
Main site UI


localhost:8501
Apps
Moodle
Portal
Portal3
Outlook
DeCuongTomTat_C...
CTDT_CNTT_CLC_K2...
Free Grammar Chec...
Visual Paradigm Onl...
MATLAB–Python–Ju...
Deploy
main
Gemini Demo
GPT-3 Demo
Visualize with GPT-3
Select a demo above.
Welcome to DataHub! 👋
#About
Group: 05
Topic: ML7 - A triple of LLMs: GPT-3, Megatron-Turing NLG, and Gemini
Description: Utilize large language models (LLMs) like GPT-3, Megatron-Turing NLG, and Gemini in building a robust question-answering system to perform data chat, data analysis and data visualization.
#Usage
Build a question-answering system that can process natural language queries related to data, extract relevant information from datasets, and provide clear answers and insights to users.
Use the LLMs for text analysis tasks like sentiment analysis, anomaly detection, and identifying unusual patterns in financial news, social media mentions, and other textual data sources.
Combine the insights from the LLMs with visualizations to create interactive dashboards for data-driven decision-making.
#Check out for demos
[Data analysis system with Gemini]
[Question-Answering system with GPT-3]
[Data Visualization with GPT-3]
👈 Select a demo from the sidebar!
Search
ENG
US
7:49 PM
3/22/2024

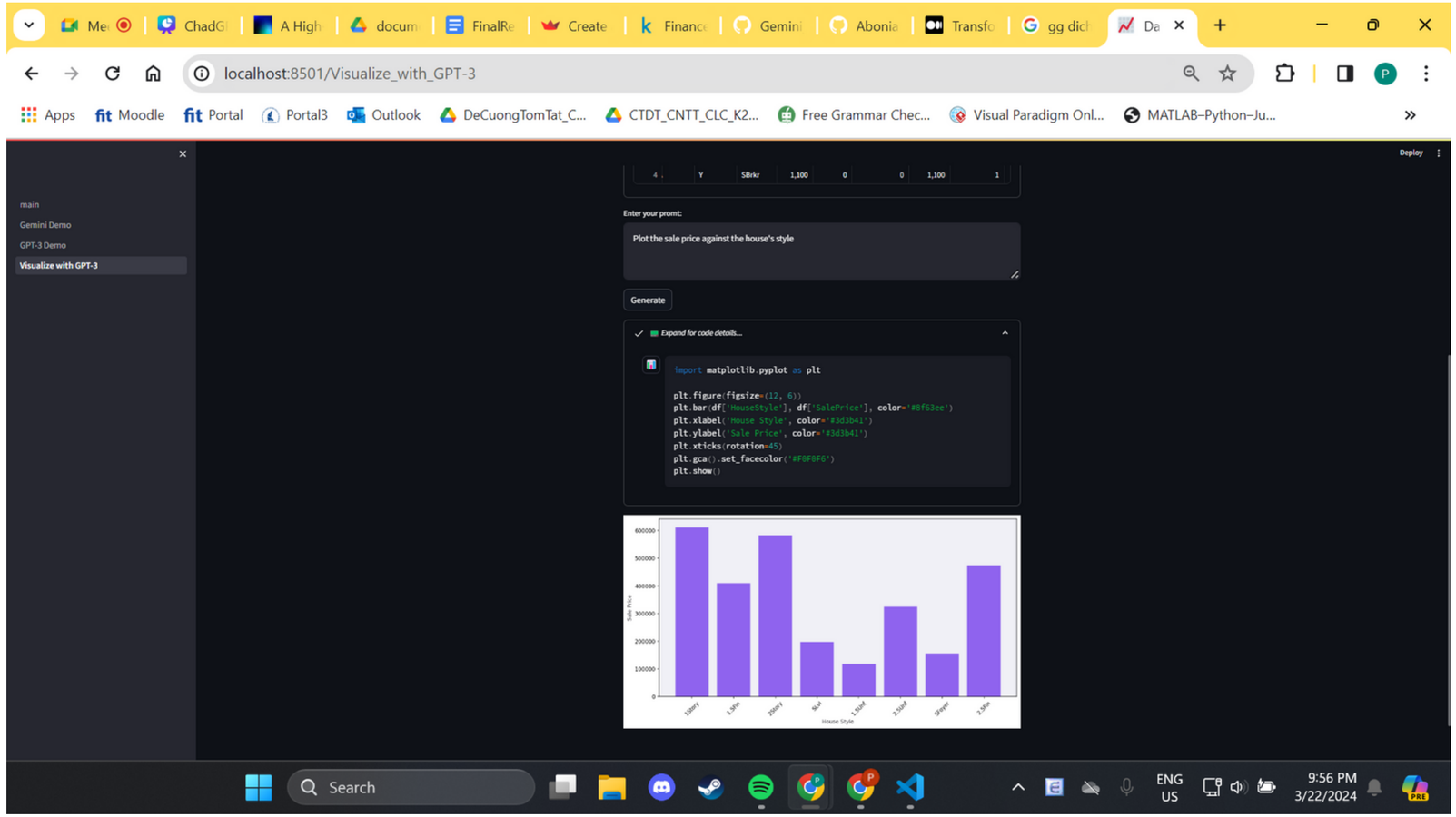
Data Analysis site UI



Data Chat site UI



Data Visualization site UI



DEMO

Discussion & Conclusion

- Are there any ways to know if the answer from these tools is correct or not?
- Are results consistent?
- Are the results secure?
- Future works?
- Is it ethical to leave data-driven decision-making to be made by probability maps?
- Its applications in academic and/or industry activities
- Its popularity (i.e., how many users choose this solution)

Q
A

VNUHCM - University of Science

THANK YOU

PRESENTATION - 2024

Thank you lecturers and everyone for listening to our presentation