

**Department of Mathematics**  
**Indian Institute of Technology Kharagpur**

**Masters Thesis Project 1**

# **Privacy Enhanced Conversational AI Using Large Language Models**

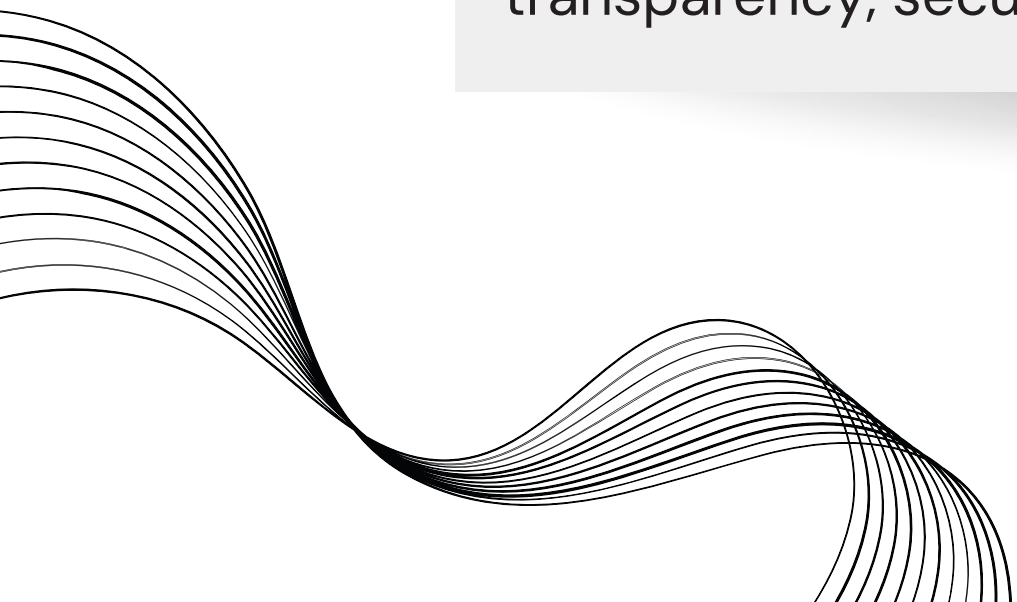
Submitted By:  
**Keerti P.Charantimath**  
19MA20059

Guided By:  
**Professor Hari Shankar Mahato**  
Department of Mathematics  
Indian Institute of Technology Kharagpur

# Introduction

**Large Language Models (LLMs)** are AI models trained on massive text data using deep learning techniques like Transformers, equipped with billions of parameters. They learn language patterns from diverse sources and excel in various natural language processing tasks, from text generation and translation to summarization and question answering.

**Blockchain technology** creates a secure system by creating a decentralized ledger spread across multiple participants, forming an immutable chain of interconnected and cryptographically linked blocks. This makes it tough to change or retrieve any data without proper authentication thus ensuring transparency, security, and data verifiability without intermediaries.



# Motivation

Combining these two powerful technologies we can achieve

**Trust and Security**  
**Streamlined Convenience**  
**Expanded Potential**

This can help us create the following technologies :

**1**

**Increasing the capabilities of  
voice-activated chatbots**

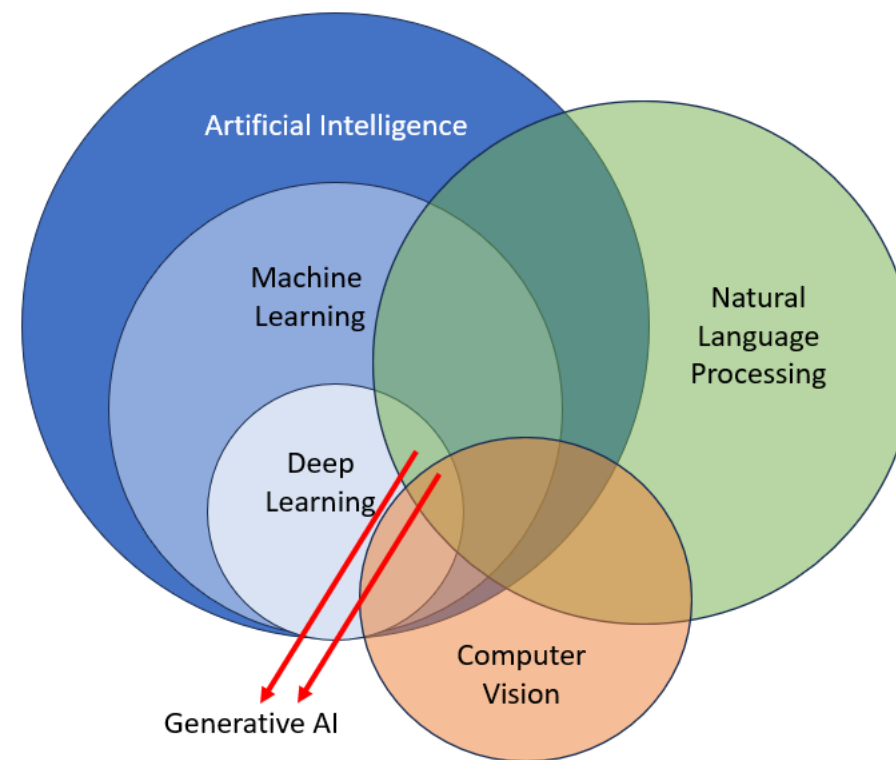
**2**

**Chatbots with tailored functionalities  
to provide financial advice**

**3**

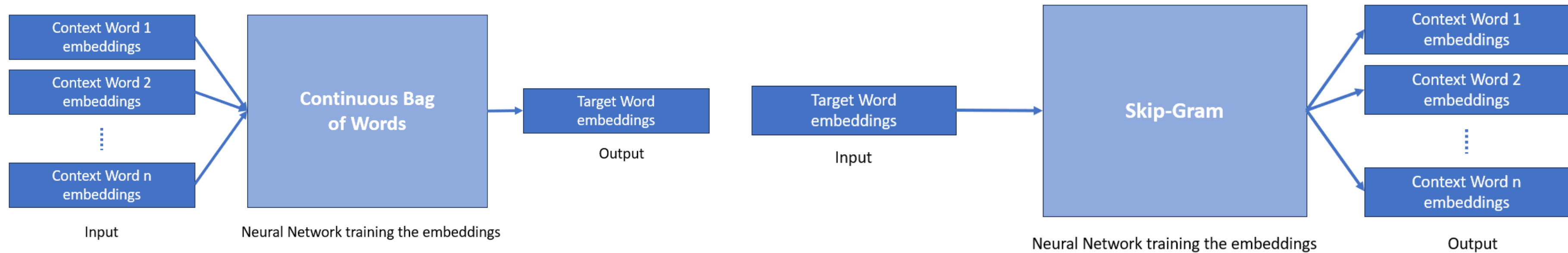
**Chatbots to guide through  
ERP or Databases**

# What Are Large Language Models?



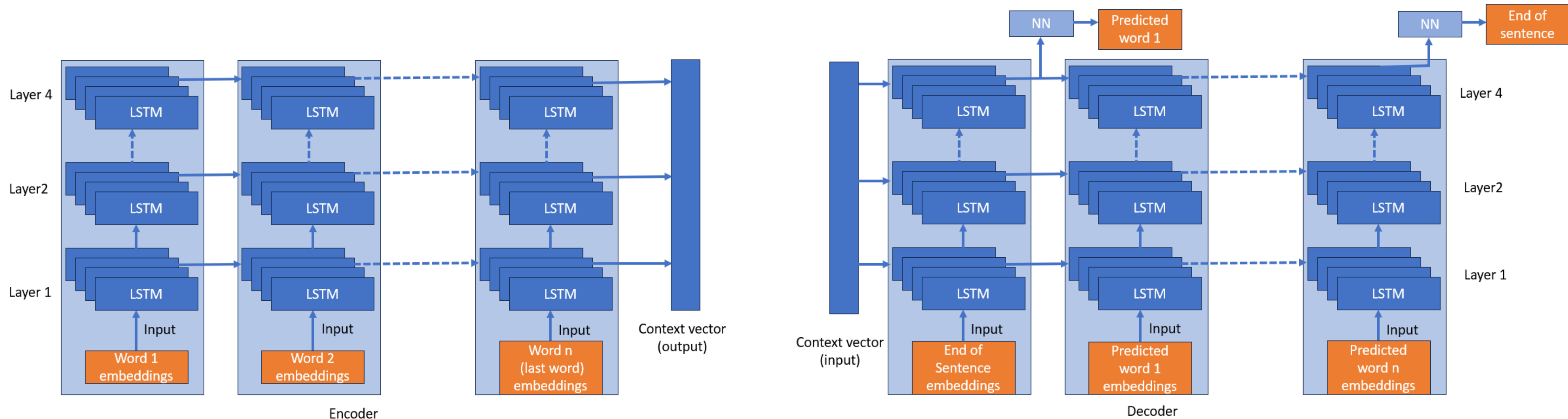
**LLMs, or Large Language Models,** are sophisticated artificial intelligence models trained on vast amounts of text data. These models, such as GPT-3, utilize deep learning techniques like Transformers and possess billions of parameters, enabling them to understand, process, and generate human-like text across various natural language processing tasks.

# Word to Vector Encoding: Word2Vec



**Word to Vector Encoding** is a technique used in natural language processing (NLP) to convert words or phrases into numerical vectors. It's based on the idea that words with similar meanings tend to occur in similar contexts within a large corpus of text.

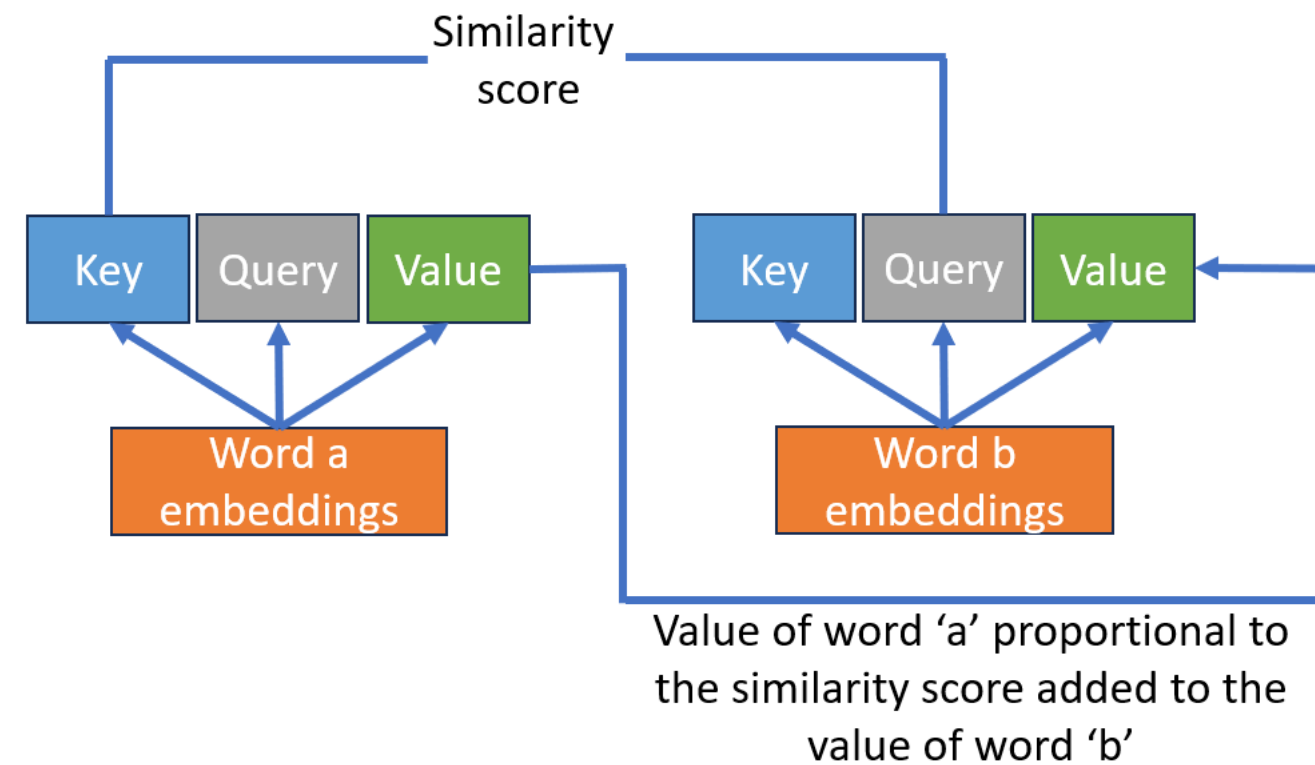
# Sequence to Sequence Model: Seq2Seq



**The Sequence-to-Sequence (Seq2Seq)** model is a neural network design employed for sequence-related tasks like language translation, text summarization, and question answering. Comprising an encoder and a decoder, it's a fundamental architecture in natural language processing (NLP) that facilitates transforming sequences from one domain to another, allowing for effective handling of sequential data.

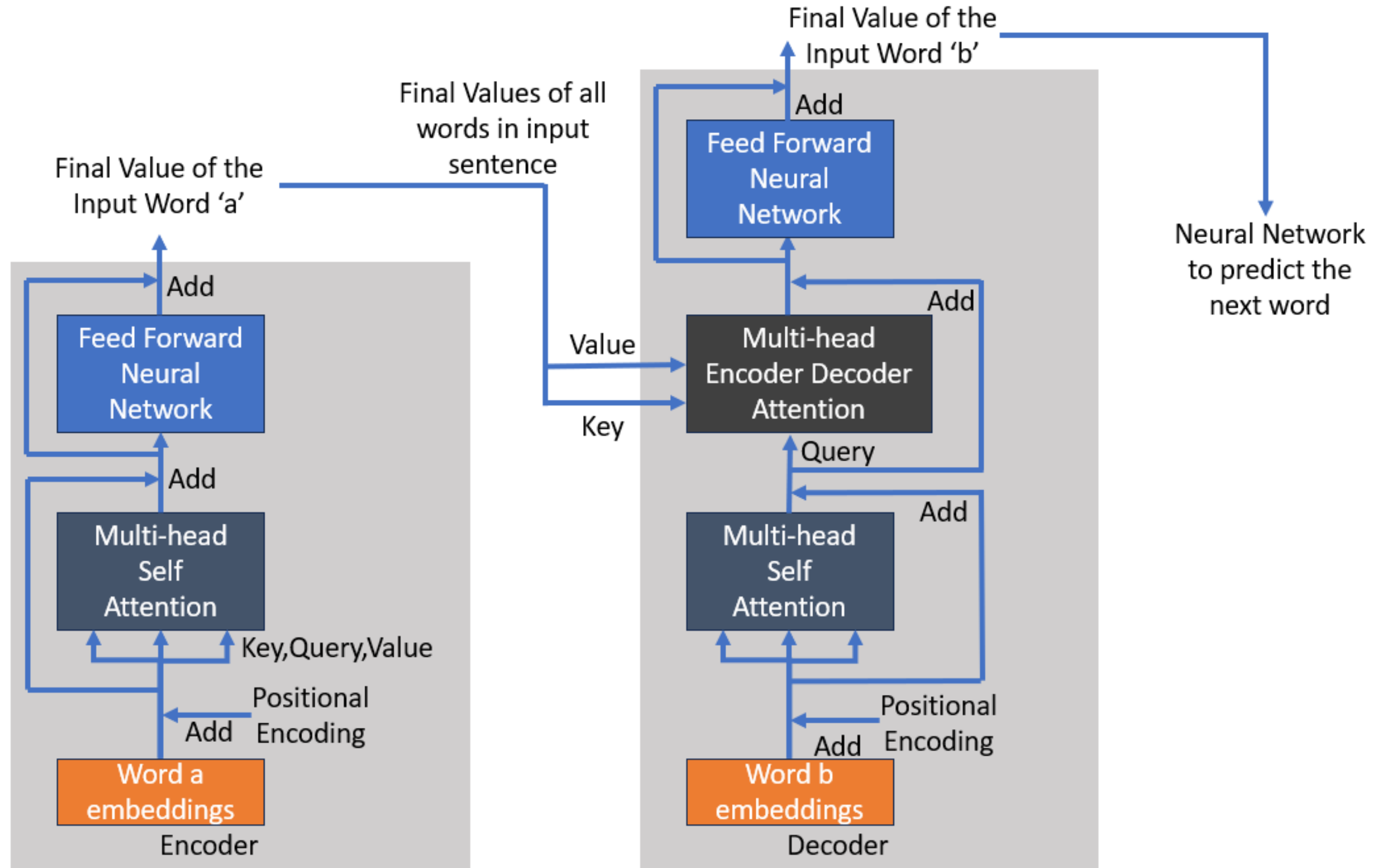


# Attention



**Attention** enables the model to focus on specific parts of the input text while processing and generating output. It involves calculating attention scores, which are similarity scores between the word being processed and all the words that would potentially have an influence on it.

# Transformer





# Types of Transformers

1

## Encoder-Decoder Transformers

- Text-to-Text Transfer Transformer
- BART - Bidirectional and Auto-Regressive Transformer

2

## Encoder-only Transformers

- BERT - Bidirectional Encoder Representations from Transformers

3

## Decoder-only Transformer

- Generative Pre-trained Transformer Series (GPT-3, GPT-2)

# Implementation

## Implementation of LLMs on local system

**Aim:** Run and interact with pre-trained Large Language Models (LLMs) on a local system, without the utilization of online LLM APIs.

To achieve the same, these steps were followed :

- **Selecting a Pre-trained Model**
- **Initializing the Model**
- **Tokenization and Input Handling**
- **Inference and Generation**

# Implementation

## Fine-tuning the LLM

**Aim:** Create an LLM model that could read, understand and answer questions based on custom personal files.

### Step 1: Selection of an LLM Model

Has better contextual understanding

LLM	CPU/GPU	Response Time (in sec)
Vicuna-7	CPU	84.983
	T4-GPU	crashed
	V100-GPU	crashed
	A100-GPU	1.74
Llama-2	CPU	86.923
	T4-GPU	3.85
	V100-GPU	3.26
	A100-GPU	3.19
Guanaco-7	CPU	78.18
	T4-GPU	crashed
	V100-GPU	crashed
	A100-GPU	5.19

Fine-tuned to maximize performance specifically on conversational tasks.

# Implementation

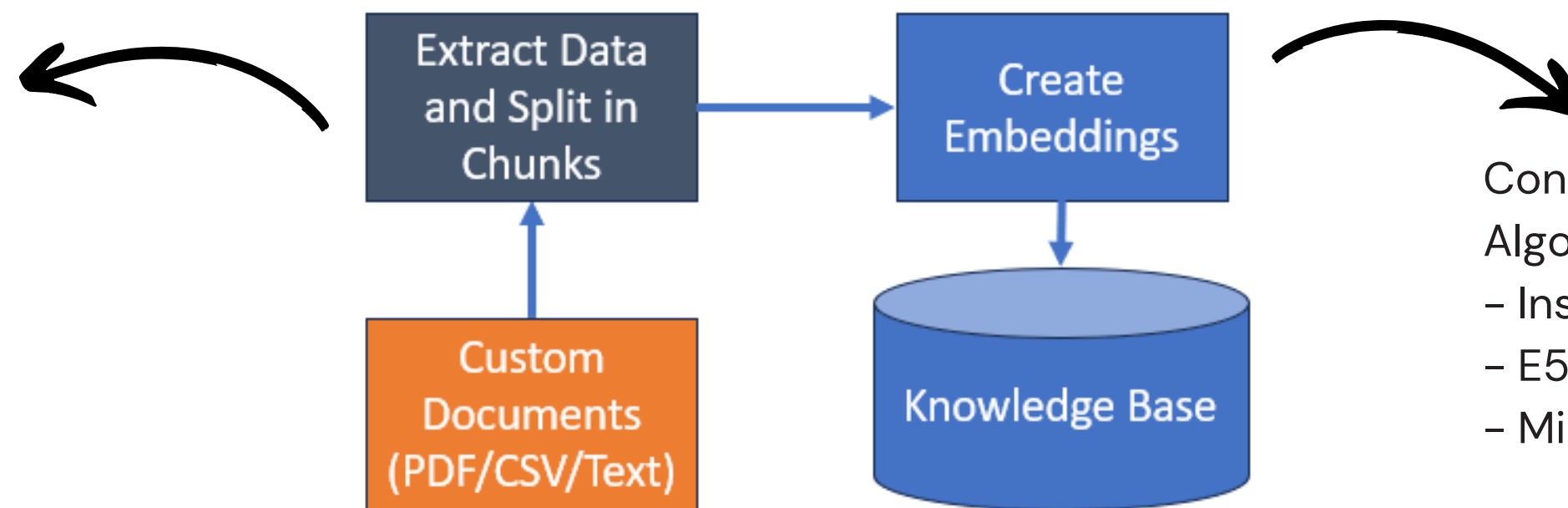
## Fine-tuning the LLM

### Step 2: Ingestion of data from personal files

Considered Chunking Strategies:

- Fixed-Size Chunking
- Recursive Chunking
- Sentence Splitting

Chunk Size- 1000  
Overlap Size- 20

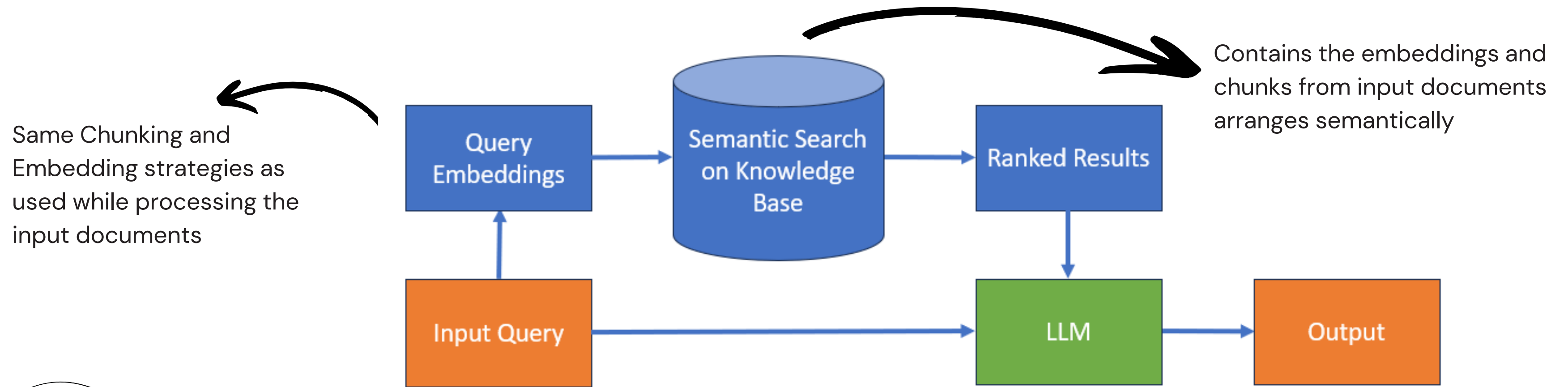


Considered Embedding Creation Algorithms:

- Instructor Algorithm
- E5 Algorithm
- MiniLM Algorithm

# Implementation

## Running LLM on Custom Files





# Results

> Question:

Keerti belongs to which country

> Answer:

Based on the provided context, Keerti P. Charantimath belongs to India.

> Question:

What are her education details

> Answer:

As an assistant, I can provide you with Keerti P. Charantimath's education details based on the provided context.

Here are the details of Keerti's education:

- \* Degree: Integrated M.Sc with B.Sc Hons.
- \* Institution: Indian Institute of Technology Kharagpur
- \* Year of completion: 2024
- \* CGPA/Percentage: 9.23/10.00

> Question:

where did keerti do her last internship

> Answer:

Based on the provided context, Keerti P. Charantimath did her last internship at Microsoft as a Software Engineering Intern.

> Question:

Where all has keerti studied?

> Answer:

Based on the provided context, Keerti P. Charantimath has studied at the following institutions:

- \* Indian Institute of Technology Kharagpur (IIT Kharagpur)
- \* MGVM's PU College, Belgaum (for her Class 12 board exams)
- \* KLE's International School, Belgaum (for her Class 10 board exams)

She has also participated in various cultural and art competitions at IIT Kharagpur, including the Inter-IIT Cultural Meet,

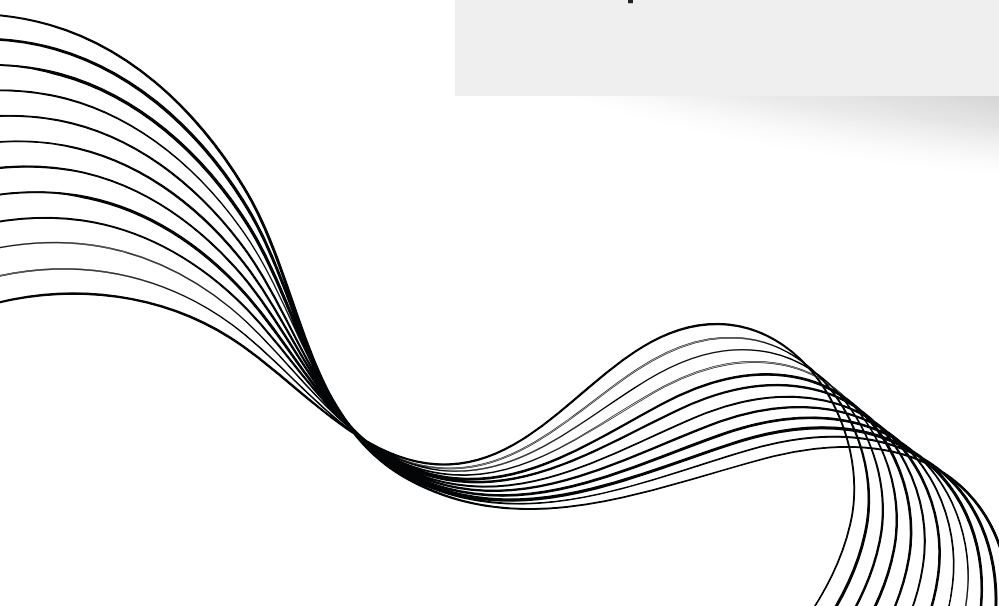


# Future Work



**Integration of LLM with Blockchain:** I plan on integrating blockchain's security features into the LLM-based chatbot system. This step aims to fortify the chatbot's framework by leveraging blockchain technology, ensuring enhanced privacy, data integrity, and trustworthiness in interactions.

**Enhancing Language Model Precision:** Additionally, I plan to refine the language model's precision and performance. By employing advanced LLM methodologies, the goal is to elevate the chatbot's language comprehension abilities, and create an effective conversational AI system.



The background features four sets of thin, black, wavy lines in the corners. The top-left and bottom-right sets are more complex, with multiple lines overlapping to create a grid-like pattern. The top-right and bottom-left sets consist of several parallel, flowing lines.

**Thank You**