# Tsinghua University Certificate Program on "Innovation & Entrepreneurship 2024

**IEDE Program Spring 2024**

**March 27, 2024**

Title of Project:

# Chatbot Implementation Using Large Language Models (Generative Pre-trained Transformer)

# PROJECT GROUP MEMBERS
# Team Name: FutureGPT

Name : Chandan Kumar Sah

University : Beihang University

Major: Software Engineering

Degree Type: Postgraduate

Member Type : Core Member

(Team Leader)

IEDE ID: 20245155

Attendance: 10/10

Work Resp: Programming, Implementation, Report, PPT

Name : Shah Syed Shazaib
University : Beihang University
Major: Power Engineering
Degree Type: Postgraduate
Member Type : Core Member
(Technical Team Leader)

IEDE ID: 2024202
Attendance: 08/10

Work Resp: Research, Data Report, PPT, Project Management

Name : Yanur Wahyu Sari Budi Asih
University: Beihang University
Major: Applied Linguistics
Degree Type: Postgraduate
Member Type : Core Member

IEDE ID: 2024201
Attendance: 08/10

Work Resp: Research, Report, PPT, Maintenance

# FutureGPT Team Members
# March 27, 2024

Tsinghua University Certificate Program on "Innovation & Entrepreneurship 2024

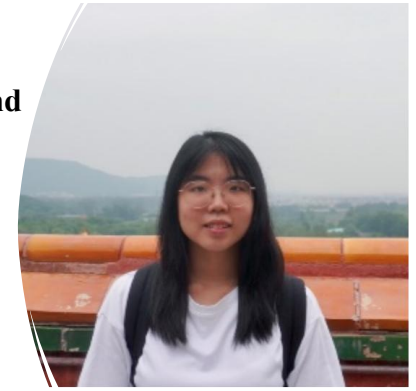
Name : S M Tarikul Islam
University: Beihang University
Major: Aeronautical Engineering
Degree Type: Postgraduate
Member Type : Core Member

IEDE ID: 2024156
Attendance: 08/10

Work Resp: Research, Report, PPT, Design

Name : Pitchanuj Chirakorn
University: Beihang University
Major: Management Science and Engineering
Degree Type: Postgraduate
Member Type : Core Member

IEDE ID: 2024153
Attendance: 09/10



Work Resp: Research, Report, PPT, Collaboration and Communication

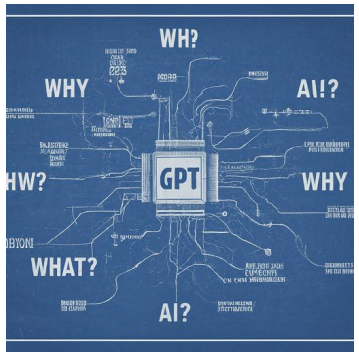
Name : Erine Wijaya Santoso
University : Beihang University
Major: Management Science and Engineering Degree
Type: Postgraduate Member
Type : Core Member

IEDE ID:2024079
Attendance: 09/10

Work Resp: Research, Report, PPT, Updates

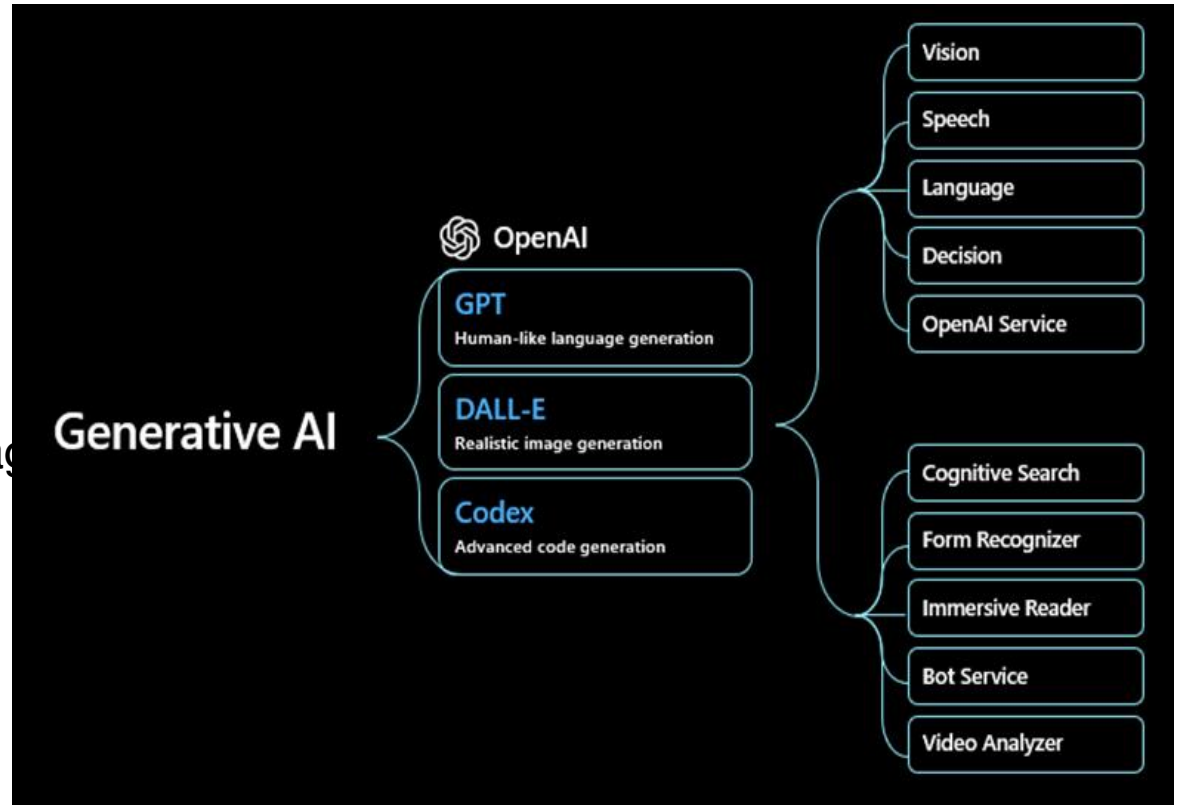
- **Contents**

# I. Introduction




WHY AI? HOW AI? WHAT AI?

- ❑ The Rise of ChatGPT
- ❑ Understanding GPT
- ❑ Building FutureGPT
- ❑ Character-Level Langua
- ❑ Model




Generative AI refers to artificial intelligence systems or models that have the capability to create new content, such as images, text, or even music, that is original and resembles human-created content.

# I. Introduction

**LLM**

- Large Language Models are a type of AI system that works with language.
- In the simplest of terms, LLMs are next-word prediction engines.
- Examples:
  - OpenAI's GPT-4
  - Google's PaLM
  - Meta's LLaMA
  - Hugging Face - Bloom

**Foundational Models**

"LLMs" specifically refers to language-focused systems, while "foundation model" is attempting to stake out a broader function-based concept, which could stretch to accommodate new types of systems in the future. (Stanford University)

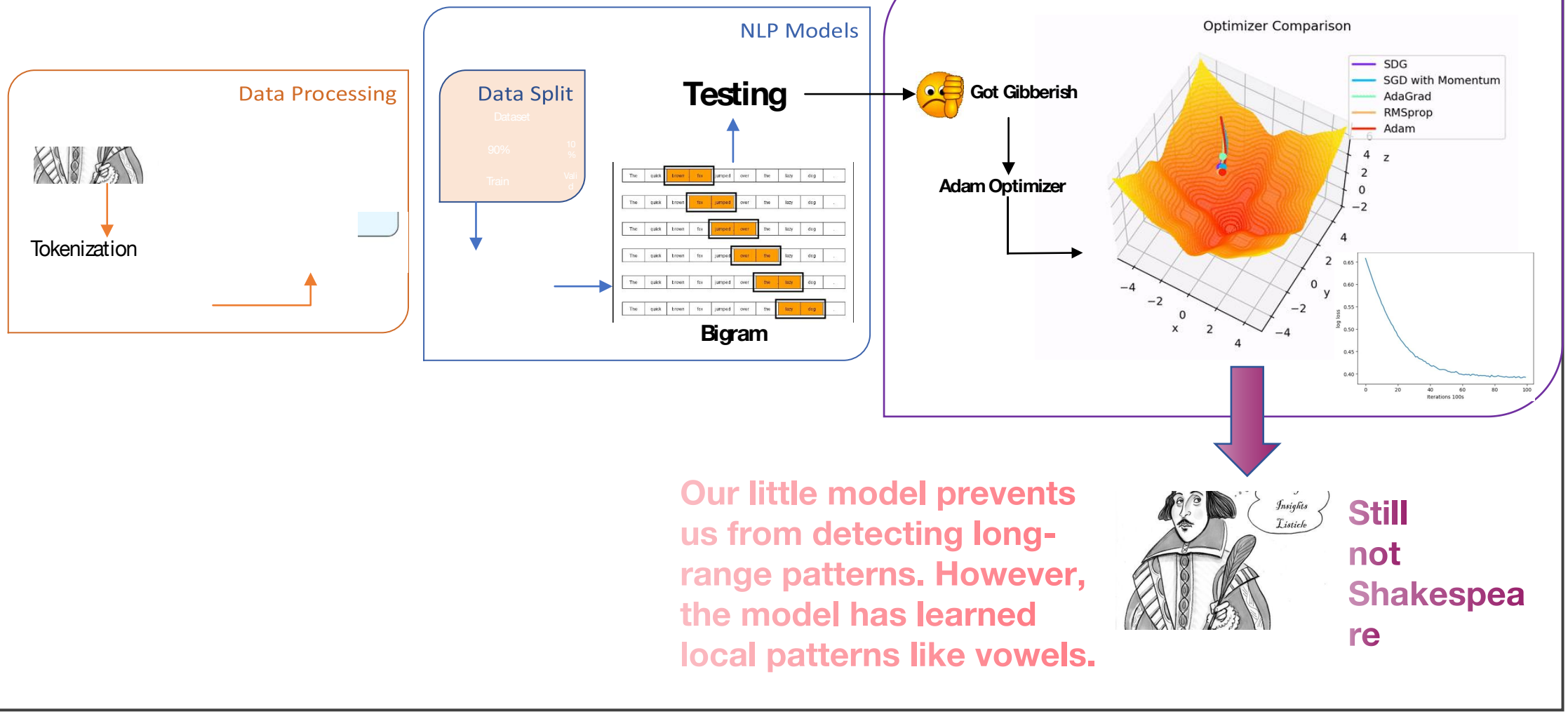
**AI Driven Chat Bots**

- UX for LLMs
- Chat GPT stands for chatbot generative pre-trained transformer
- They have LLMs behind them
- Use prompts for conversation
- Examples:
  - Open AI Chat GPT
  - Google BARD (multi modal)

**Fine Tunning**

- To use LLMs you need to fine tuning and distillation
- Fine Tuning Examples:
  - Reinforcement Learning with Human Feedback (Open AI)
  - Active Learning (UiPath)

# II. Methodology



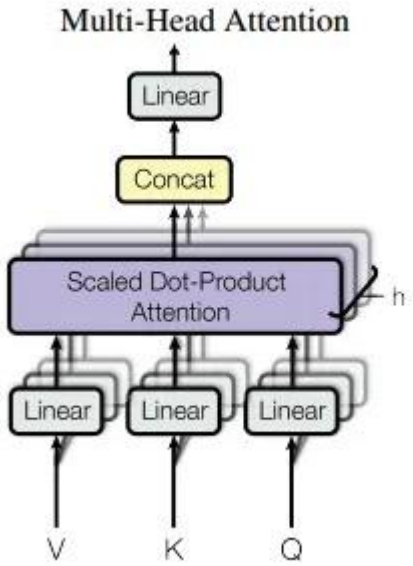

**Our little model prevents us from detecting long-range patterns. However, the model has learned local patterns like vowels.**




**Still not Shakespeare**

# III. Research & Implementation

- Transformer Model Architecture:
- Encoder-Decoder Structure:
- Attention Mechanism:
- Multi-Head Attention:



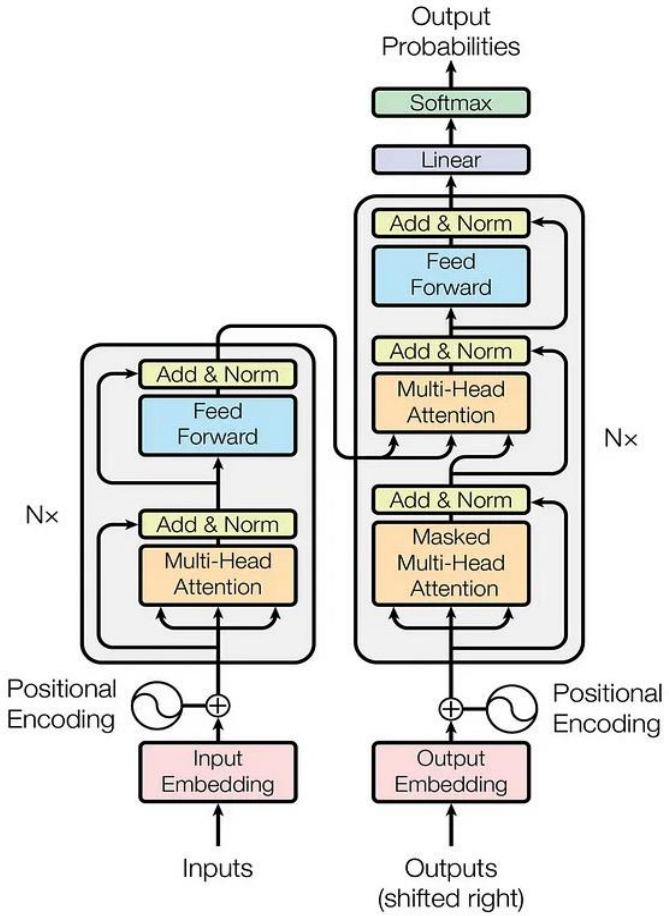



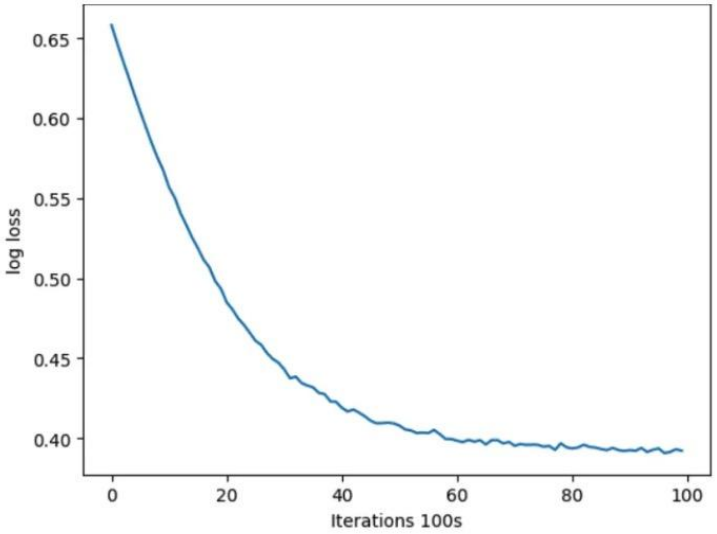

Figure 1: The Transformer - model architecture.

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)$$

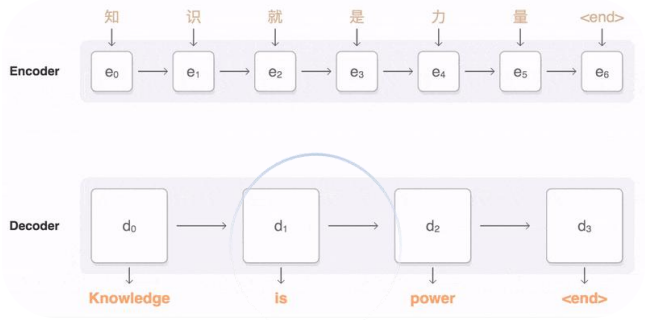
- Attention is a communication mechanism with a directed graph where nodes aggregate information via weighted sums in a data-dependent manner.
- Positional encoding is used to encode node positions, as attention does not have a notion of space, unlike convolutional mechanisms.
- Examples across the batch dimension do not communicate with each other.
- Encoder blocks enable communication between all nodes, useful for tasks like sentiment analysis.
- Self-attention is used when keys, queries, and values come from the same source, while cross-attention involves keys and values from different sets of nodes.

```
when input is tensor([18]) the target: 47
when input is tensor([18, 47]) the target: 56
when input is tensor([18, 47, 56]) the target: 57
when input is tensor([18, 47, 56, 57]) the target: 58
when input is tensor([18, 47, 56, 57, 58]) the target: 1
when input is tensor([18, 47, 56, 57, 58,  1]) the target: 15
when input is tensor([18, 47, 56, 57, 58,  1, 15]) the target: 47
when input is tensor([18, 47, 56, 57, 58,  1, 15, 47]) the target: 58
```

- The transformer is trained to handle variable length contexts, ranging from small contexts of size 1 to larger contexts of size 8.
- This capability is particularly advantageous during inference time, allowing the model to adapt to different input lengths dynamically.

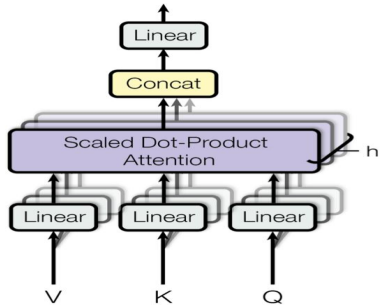


# IV. Tools Used Conclusion

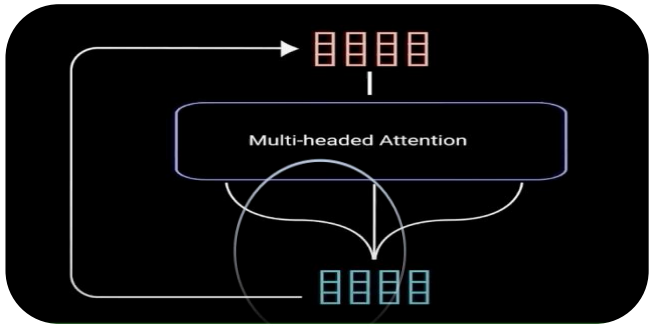



Attention Mechanism

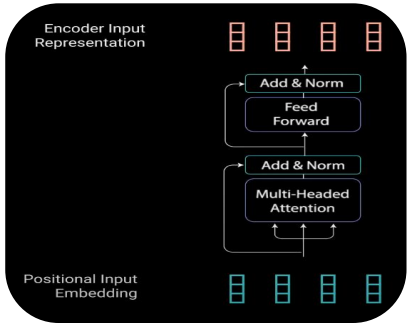

Bigram Model



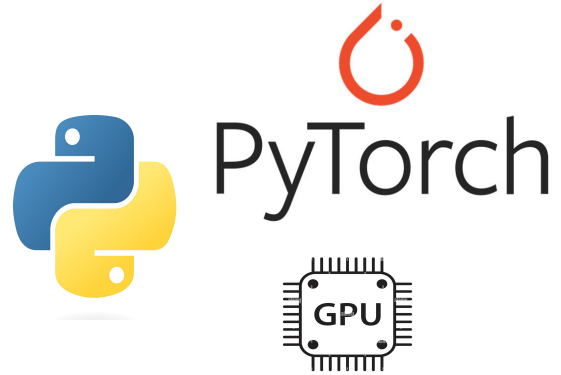

Multi-Head Attention




Feed-forward with attention




Multi-Head Attention

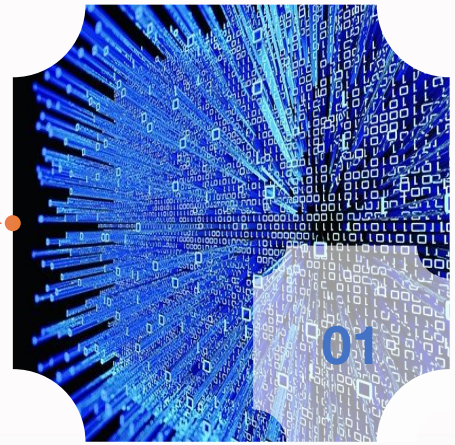



Software and Hardware Tools

# **V.** Future Plan and Recommendation

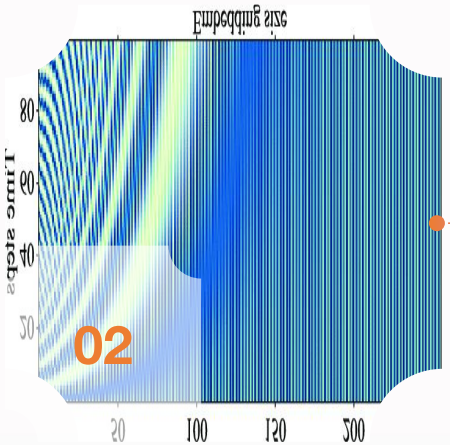
## Increasing Model Size

Experimenting with larger embedding dimensions and more attention heads can improve the model's capacity to learn complex patterns.




## Different positional Encoding

While this project used a simple positional embedding table, other methods like sinusoidal encodings can be explored.

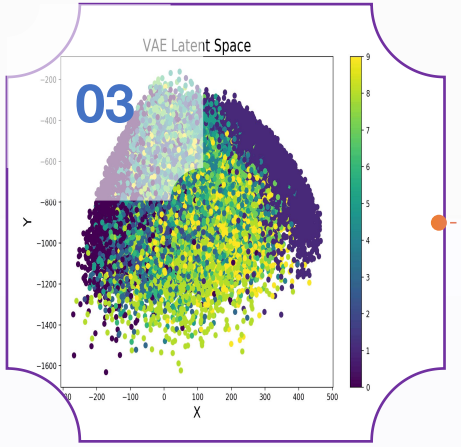



## Adding More Layers

Stacking multiple self-attention and feed-forward layers can allow the model to capture deeper relationships within the data.


## Different Training Stratagies

Techniques like gradient clipping and learning rate scheduling can be used to optimize the training process.



# V. Summary

**“Although the study wasn’t able to achieve Shakespeare due to resource restriction, the result did show significant abilities to capture contextual patterns and hint towards much higher performance if expaned further with proper resource utilisation”**

**01 Shakespeare**

Utilized small Shakespeare dataset for character-level encoding.

**02 Training and validation**

Data spliting

**03 GPT -- Bigram**

Click here to add the text, and please try to explain your point of view as succinctly as possible.

**05 Self-Attention**

**Multi-Head Attention**

Increase the word forecast accuracy by weighing the contextual meaning

**06 Feed-Forward Layer**

Added a feed-forward layer for complex information processing.

# Thanks