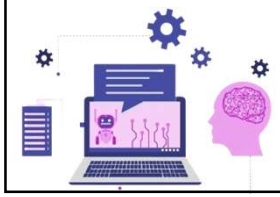
Natural Language Processing And Generative AI



Dr. Shailesh Sivan
Principal AI Architect, Laennec AI
Assistant Professor, CUSAT(on leave)
shaileshsivan@gmail.com

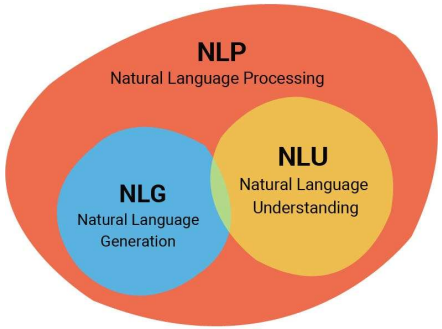
What is NLP??

Natural language processing (NLP)

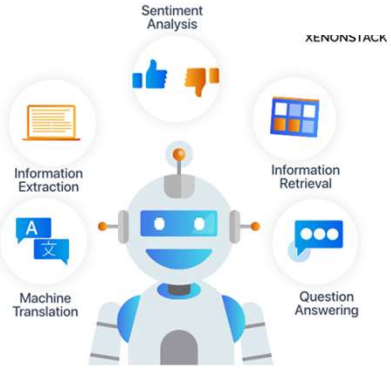
- A branch of Artificial Intelligence that gives machines the ability to understand natural human language.

billions of text data being generated every day and most of them are unstructured.

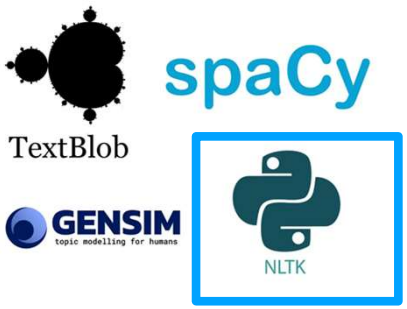
What is NLP??



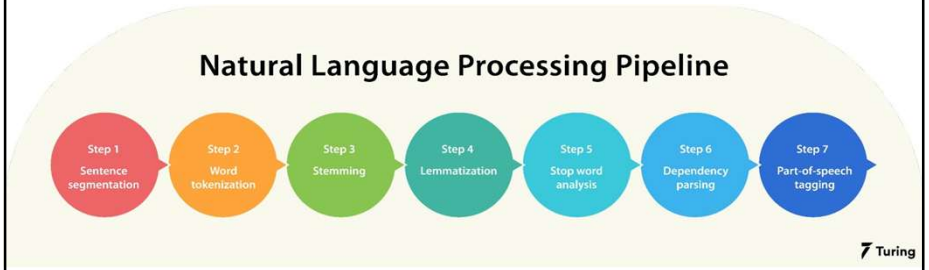
Applications of Natural Language Processing in Different Domains



Python Libraries for NLP

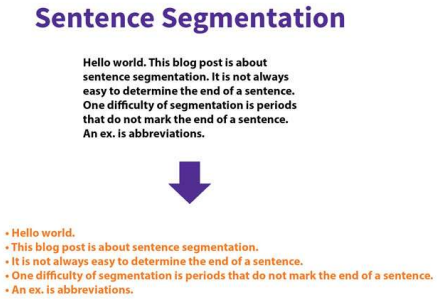


Common Terminologies



Sentence Segmentation

Divides the entire paragraph into different sentences for better understanding.



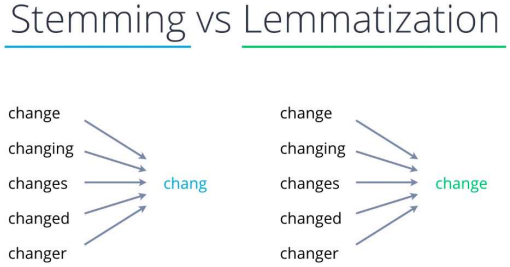
Tokenization

breaks the sentence into separate words or tokens to understand the context of the text.

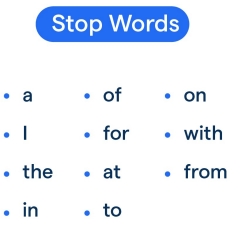
Stemming and Lemmatization

Stemming
- cuts off prefixes and suffixes to reduce a word to its root form called the stem, may not be a valid word in the language

Lemmatization
- Reducing words to their base or dictionary form, known as the lemma.



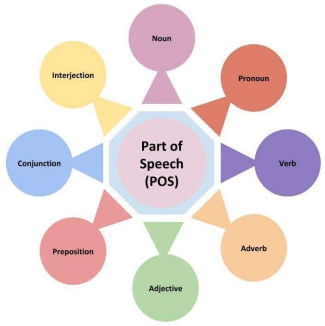
Stop Words

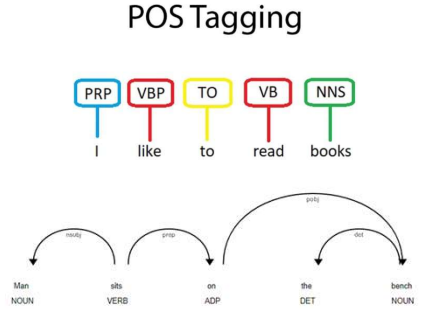


When was the first computer invented?
How do I install a hard disk drive?
How do I use Adobe Photoshop?
Where can I learn more about computers?
How to download a video from YouTube
What is a special character?
How do I clear my Internet browser history?
How do you split the screen in Windows?
How do I remove the keys on a keyboard?
How do I install a hard disk drive?

words that are so widely used that they carry very little useful information.

Part Of Speech(POS)





Label words in a text with their grammatical category, such as noun, verb, or adjective

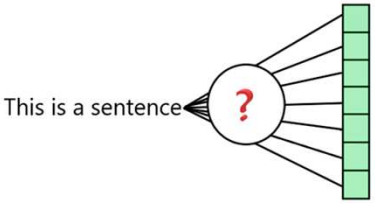
Named Entity Recognition

contentSkip to site indexPoliticsSubscribeLog InSubscribeLog InToday's PaperAdvertisementSupported ORG byF.B.I. Agent Peter Strzok PERSON , Who Criticized Trump PERSON in Texts, Is FiredImagePeter Strzok, a top F.B.I. GPE counterintelligence agent who was taken off the special counsel investigation after his disparaging texts about President Trump PERSON were uncovered, was fired. CreditT.J. Kirkpatrick PERSON for The New York TimesBy Adam Goldman ORG and Michael S. SchmidtAug PERSON . 13 CARDINAL , 2018WASHINGTON CARDINAL — Peter Strzok PERSON , the F.B.I. GPE senior counterintelligence agent who disparaged President Trump PERSON in inflammatory text messages and helped oversee the Hillary Clinton PERSON email and Russia GPE investigations, has been fired for violating bureau policies, Mr. Strzok PERSON 's lawyer said Monday DATE .Mr. Trump and his allies seized on the texts — exchanged during the 2016 DATE campaign with a former F.B.I. GPE lawyer, Lisa Page — in PERSON assailing the Russia GPE investigation as an illegitimate "witch hunt." Mr. Strzok PERSON , who rose over 20 years DATE at the F.B.I. GPE to become one of its most experienced counterintelligence agents, was a key figure in the early months DATE of the inquiry.Along with writing the texts, Mr. Strzok PERSON was accused of sending a highly sensitive search warrant to his personal email account.The F.B.I. GPE had been under immense political pressure by Mr. Trump PERSON to dismiss Mr. Strzok PERSON , who was removed last summer DATE from the staff of the special counsel, Robert S. Mueller III PERSON . The president has repeatedly denounced Mr. Strzok PERSON in posts on

Process of identifying and classifying key information (entities) such as names of people, organizations, locations, dates, and more within unstructured text.

Text Feature Extraction and Classification

Text Features Extraction



The quick brown fox jumped over the brown dog

the	quick	brown	fox	jumped	over	the	brown	dog
1	4	13	9	5	2	1	13	23

Turning text into vectors that can be then fed to machine learning models in a classical way

Types

Machine learning algorithms cannot work with raw text directly, the text must be converted into numbers. Specifically, vectors of numbers.

- N- Grams
- Bag-of-Words
- Term Frequency -Inverse Document Frequency (TF-IDF)
- Word Embedding

N-grams

N-grams are the combination of multiple words used together.

N-grams with N=1 are called unigrams.

N=2 - bigrams,

N=3 - trigrams and so on..

This is Big Data AI Book

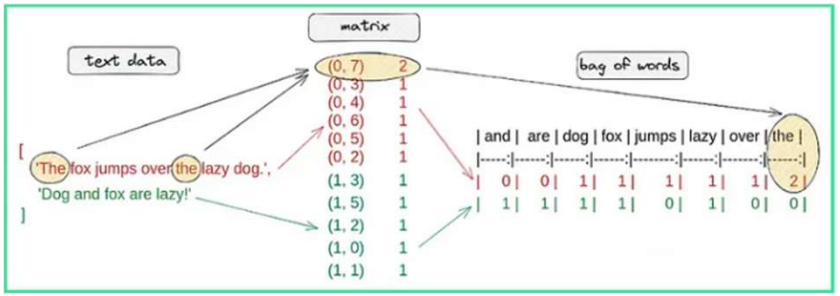
Uni-Gram	This	Is	Big	Data	AI	Book
Bi-Gram	This is	Is Big	Big Data	Data AI	AI Book	
Tri-Gram	This is Big	Is Big Data	Big Data AI	Data AI Book		

Bag of Words (BoW)

- used to analyze text and documents based on **word count.**
- model does not account for word order within a document.

	about	bird	heard	is	the	word	you
About the bird, the bird, bird bird bird	1	5	0	0	2	0	0
You heard about the bird	1	1	1	0	1	0	1
The bird is the word	0	1	0	1	2	1	0

Bag of Words (BoW)



Bag of Words (BoW) Limitation

'The sky is blue and beautiful',
'The king is old and the queen is beautiful',
'Love this beautiful blue sky',
'The beautiful queen and the old king']

	and	beautiful	blue	is	king	love	old	queen	sky	the	this
0	1	1	1	1	0	0	0	0	1	1	0
1	1	1	0	2	1	0	1	1	0	2	0
2	0	1	1	0	0	1	0	0	1	0	1
3	1	1	0	0	1	0	1	1	0	2	0

	beautiful	beautiful blue	beautiful queen	blue	blue beautiful	blue sky	king	king old	love	love beautiful	old	old king	old queen	queen	queen beautiful	queen old	sky	sky blue
0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	1
1	1	0	0	0	0	0	1	1	0	0	1	0	1	1	1	0	0	0
2	1	1	0	1	0	1	0	0	1	1	0	0	0	0	0	0	1	0
3	1	0	1	0	0	0	1	0	0	0	1	1	0	1	0	1	0	0

Term Frequency – Inverse Document Frequency Matrix

Measures the importance of words in a document

Term frequency
- The number of times a word appears in a document
- Ratio of the count of a word's occurrence in a document and the number of words in the document

Inverse Document frequency
- Measure of how common or rare a word is in the entire corpus of documents
- IDF of word t represents the ratio of number of documents in the corpus with number of documents with word t in them

$$TF(t,d) = \frac{\text{Number of times term } t \text{ occurs in document } d}{\text{Total number of terms in document } d} \quad \text{IDF}(t,d) = \log\left(\frac{\text{Total number of documents in corpus } N}{\text{Number of documents containing the term } t}\right)$$

$$\text{TFIDF}(t,d,N) = \text{TF}(t,d) \times \text{IDF}(t,d)$$

Term Document – Inverse Document Frequency Matrix

$$TF(t,d) = \frac{\text{Number of times term } t \text{ occurs in document } d}{\text{Total number of terms in document } d} \qquad IDF(t,d) = \log\left(\frac{\text{Total number of documents in corpus } N}{\text{Number of documents containing the term } t}\right)$$

$$\text{TFIDF}(t,d,N) = \text{TF}(t,d) \times \text{IDF}(t,d)$$

Actual text

I read the svm algorithm article in dataaspirant blog

I read the randomforest algorithm article in dataaspirant blog

Preprocessed text

read svm algorithm article dataaspirant blog

read randomforest algorithm article dataaspirant blog

TF-IDF Calculation Example

Words	Count		Term Frequency (TF)		Inverse Document Frequency (IDF)	TF * IDF	
	Document 1	Document 2	Document 1	Document 2		Document 1	Document 2
read	1	1	0.17	0.17	0	0	0
svm	1	0	0.17	0	0.3	0.05	0
algorithm	1	1	0.17	0.17	0	0	0
article	1	1	0.17	0.17	0	0	0
dataaspirant	1	1	0.17	0.17	0	0	0
blog	1	1	0.17	0.17	0	0	0
randomforest	0	1	0	0.17	0.3	0	0.05

Term Document – Inverse Document Frequency Matrix

Actual text

Petrol cars are cheaper than diesal cars

Diesal is cheaper than petrol

Preprocessed text

Petrol cars cheaper diesal cars

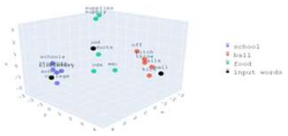
Diesal cheaper petrol

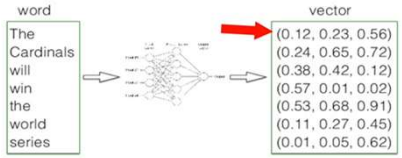
Words	TF = Frequency(word)		IDF =[$\log_e((d+1)/(df(t)+1)]+1$	TF-IDF	
	Document 1	Document 2		Document 1	Document 2
car	2	0	$\log_e(3/2)$ +1 => 1.405465083	2.8109302	0
cheaper	1	1	$\log_e(3/3)$ + 1 => 1	1	1
diesal	1	1	$\log_e(3/3)$ + 1 => 1	1	1
petrol	1	1	$\log_e(3/3)$ + 1 => 1	1	1

In the first document the term "cars" is the most relevant term as it has the highest tf-idf value
In the second document most of the terms have the same tf-idf value and have equal relevance.

Word Embedding

- A technique that represents words as numbers so that computers can process them.
- Capture semantic and syntactic relationship between texts and represent them in the form of vectors
- IDEA - Similar words will have a minimum distance between their vectors.





Word2Vec
neural approach for generating word embeddings

- **Continuous Bag of Words (CBOW).**
- **Skip-grams**

Continuous Bag of Words Vs Skip-gram

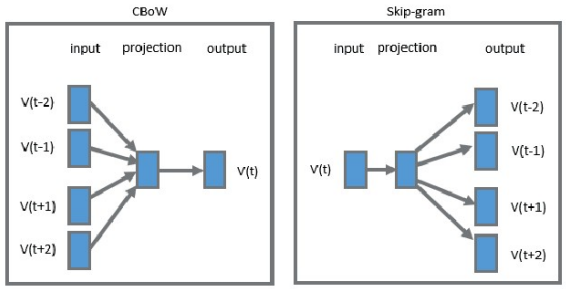
The sun is bright and the weather is pleasant.		
Windowsize = 2 (2 words to the left and 2 words to the right)		
Position	Target Word	Context Words
1	the	[sun, is]
2	sun	[the, is, bright]
3	is	[the, sun, bright, and]
4	bright	[sun, is, and, the]
5	and	[is, bright, the, weather]
6	the	[bright, and, weather, is]
7	weather	[and, the, is, pleasant]
8	is	[the, weather, pleasant]
9	pleasant	[weather, is]

Continuous Bag of Words Vs Skip-gram

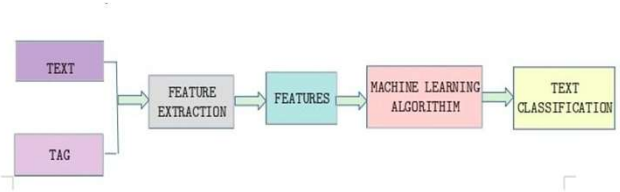
Neural network-based algorithm

CBOW : Predicts a target word given its surrounding context words.

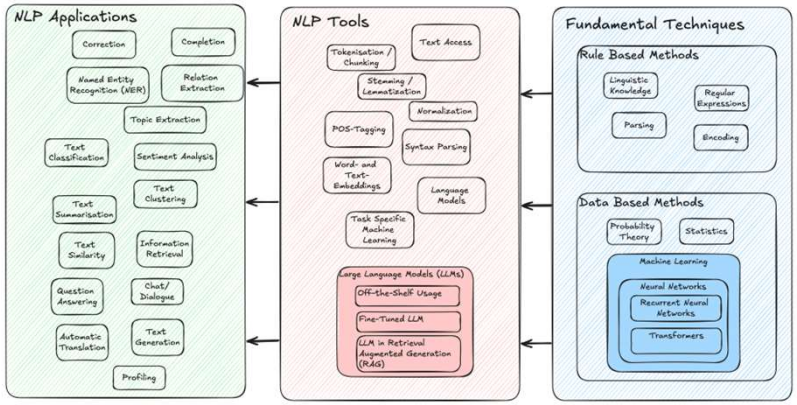
Skip-gram : Predicts surrounding words based on a specific word called the "target word."



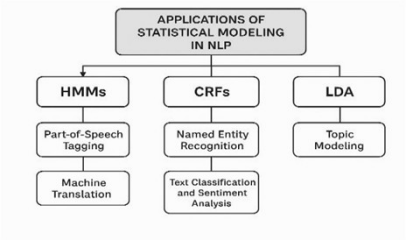
Text Classification -Pipeline

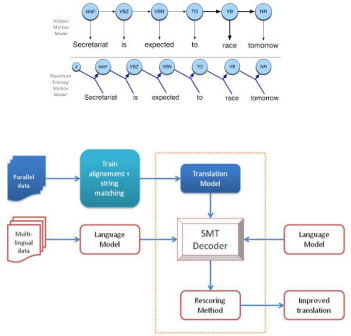


Outline Of NLP Taxonomy



Statistical NLP Models





What Is Machine Learning?

"Learning is any process by which a system improves performance from experience."

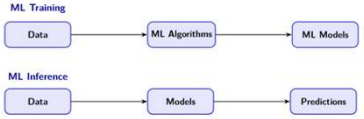
- Herbert Simon

Definition by Tom Mitchell (1998):

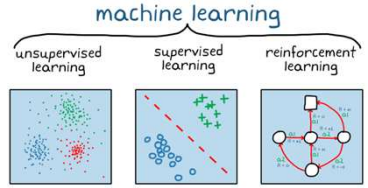
Machine Learning is the study of algorithms that
- improve their performance **P**
- at some task **T**
- with experience **E**.

A well-defined learning task is given by <P,T,E>

Machine Learning is a process where a computer uses *data* to learn ***hidden patterns*** through an ***ML algorithm***, and the output of this learning is an ***ML model*** that can make ***predictions*** on new data.



Machine Learning



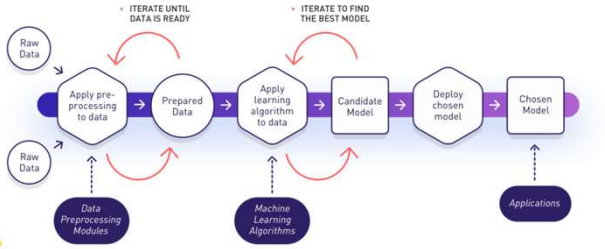
Problems with Uncertainty

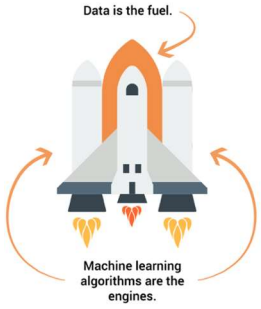
Approximating the patterns to Generalization

Data with hidden patterns

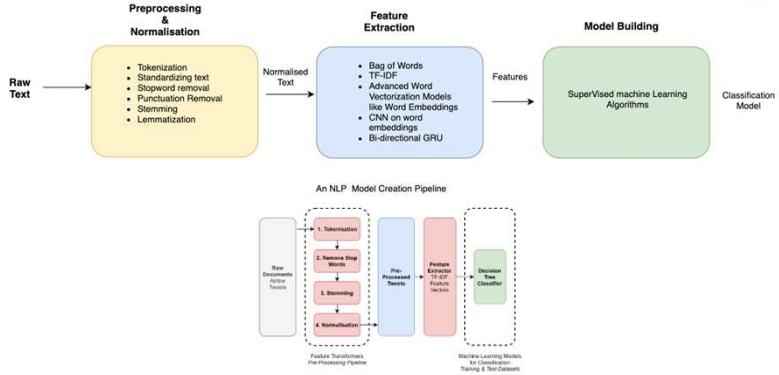
Minimum error to predict unseen data

MACHINE LEARNING PIPELINE

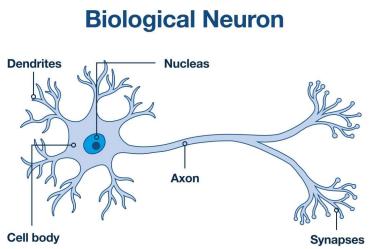


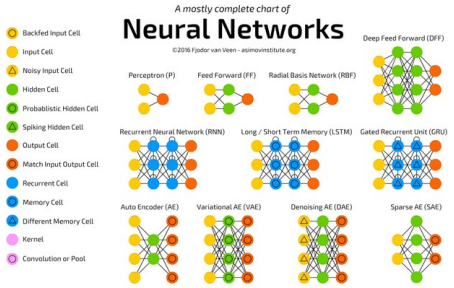


MACHINE LEARNING PIPELINE

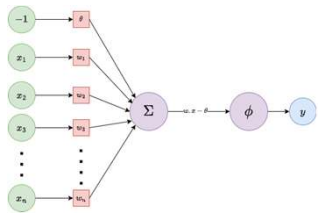


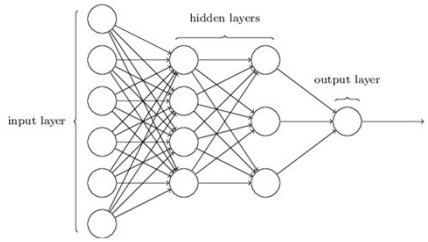
NEURAL NETWORKS : PROMINENT ML FAMILY





PERCEPTRON & MULTILAYER PERCEPTRON





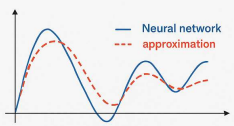
UNIVERSAL APPROXIMATION THEOREM

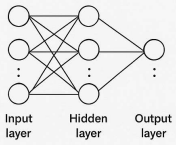
- A feedforward neural network with a single hidden layer and finite neurons can approximate any continuous function on a compact subset of $\mathbb{R}^n$, given suitable activation functions.

For any continuous function $f: \mathbb{R}^n \to \mathbb{R}$ and for any $\varepsilon > 0$, there exists a neural network f such that:

$$|f(x) - f(x)| < \varepsilon$$

for all x in the compact subset.

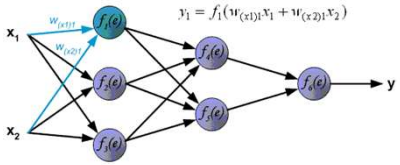




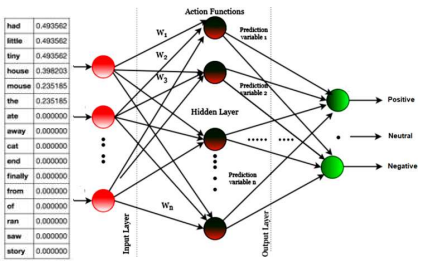
- Key Intuition:
  - ✓ Neural networks are universal function approximators
  - ✓ They can represent any continuous mapping with enough neurons in the hidden layer

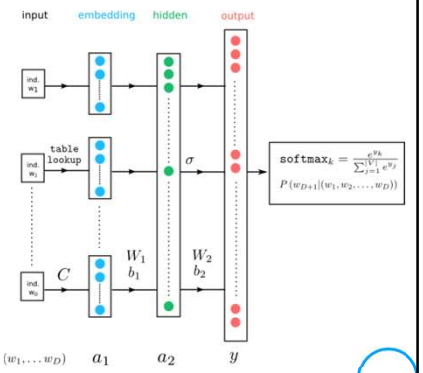
BACKPROPAGATION LEARNING ALGORITHM 'BP'

FP

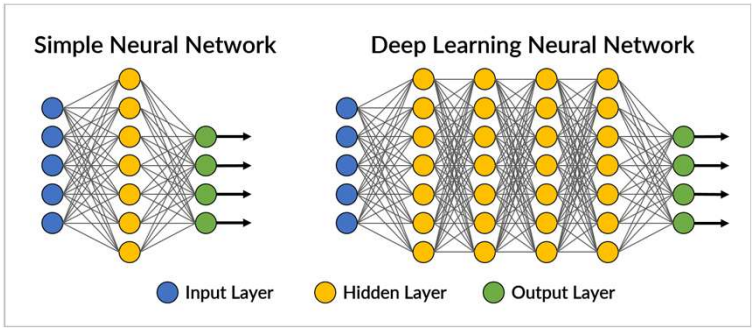


NEURAL LANGUAGE PROCESSING

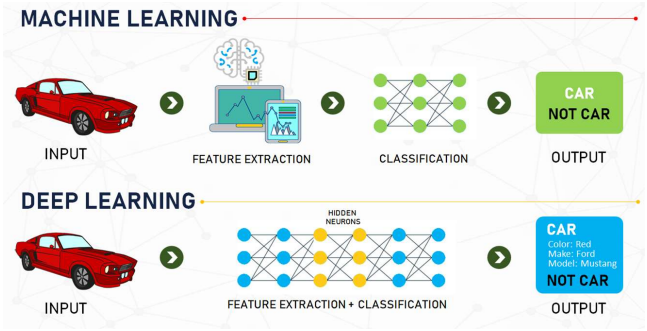




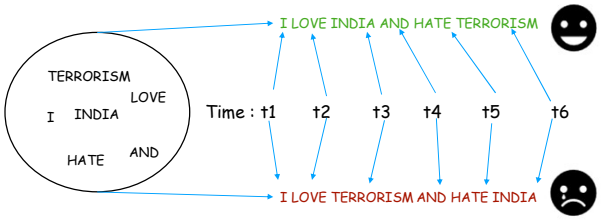
DEEP NEURAL NETWORK



DEEP LEARNING



TEMPORAL DATA

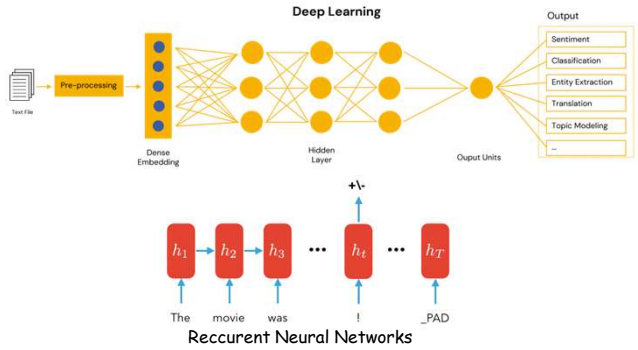


"There is nothing to like in this movie."

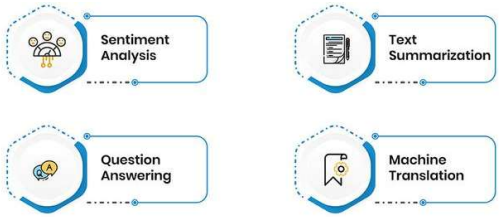
AGCCCCTGTGAGGAACTAG

COMPUTER VISION

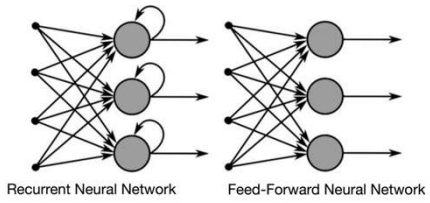
DEEP LEARNING(NLP)



DEEP LEARNING FOR NLP



Recurrent Neural Network

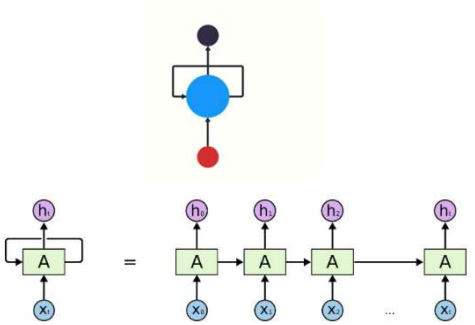


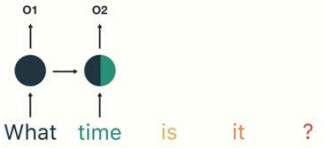
RNN the information cycles through a loop. When it makes a decision, it considers the current input and also what it has learned from the inputs it received previously.

A recurrent neural network, however, is able to remember those characters because of its internal memory. It produces output, copies that output and loops it back into the network.

Simply put: recurrent neural networks add the immediate past to the present.

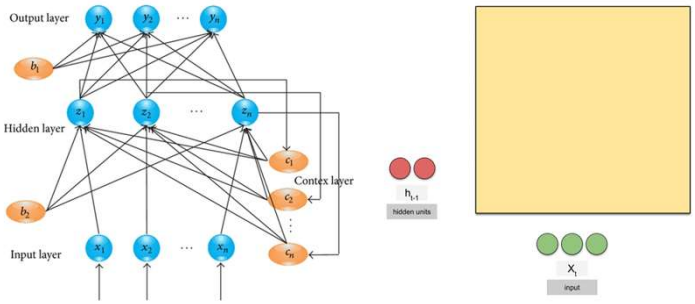
Recurrent Neural Network



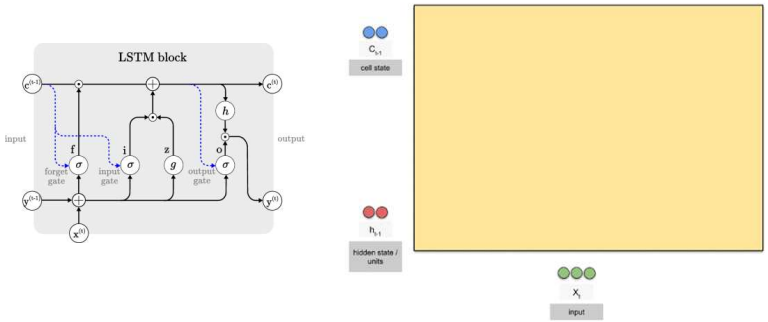


RNN has two inputs: the present and the recent past. This is important because the sequence of data contains crucial information about what is coming next, which is why a RNN can do things other algorithms can't.

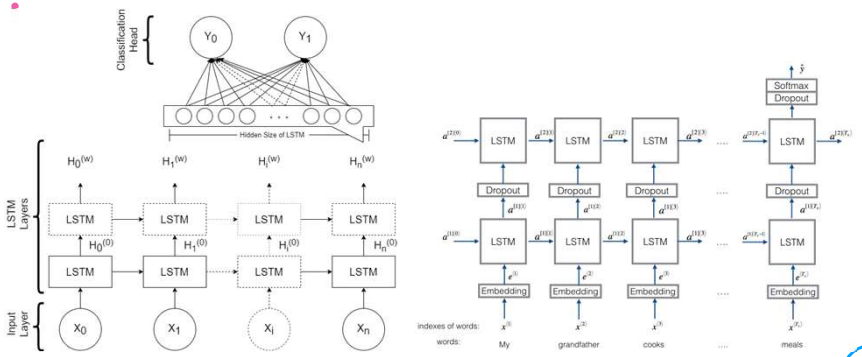
Recurrent Neural Network



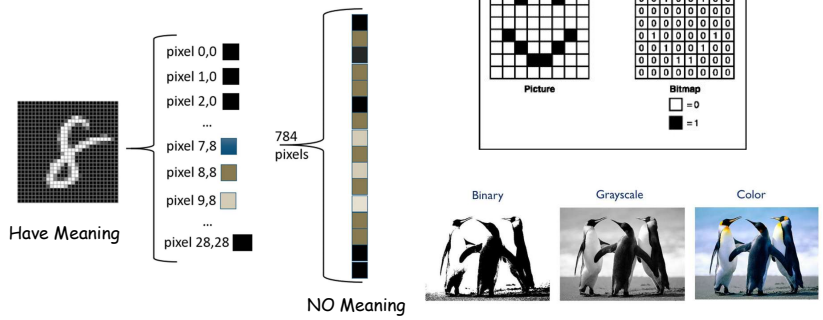
Long Short Term Memory



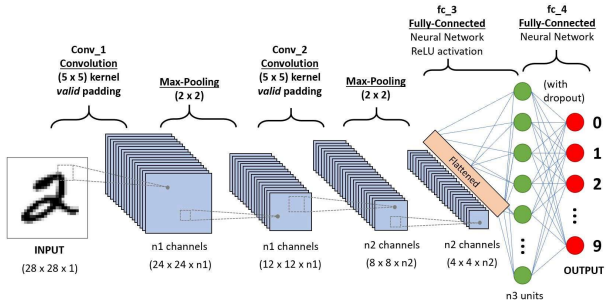
TEXT CLASSIFICATION USING RNN



SPATIAL DATA

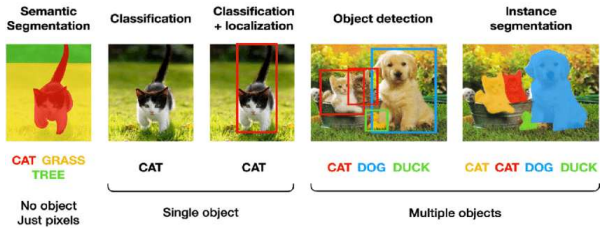


DEEP LEARNING(VISION)

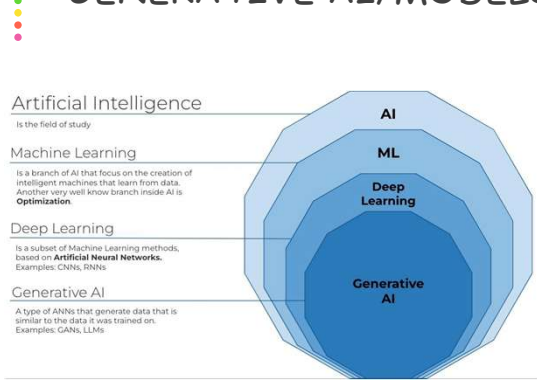


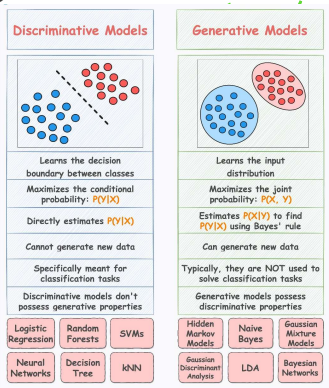
Convolutional Neural Networks

DEEP LEARNING FOR VISION

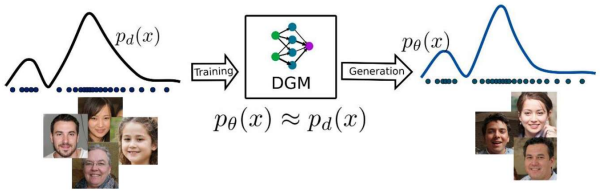


GENERATIVE AI/MODELS

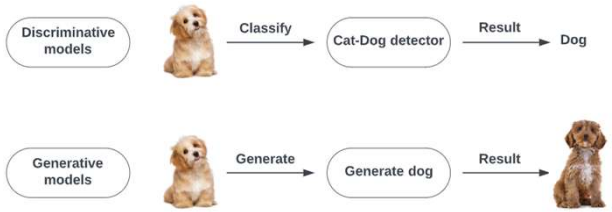




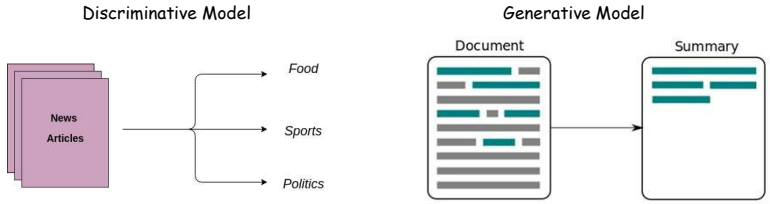
Deep Generative Models



GENERATIVE VS DISCRIMINATIVE MODELS

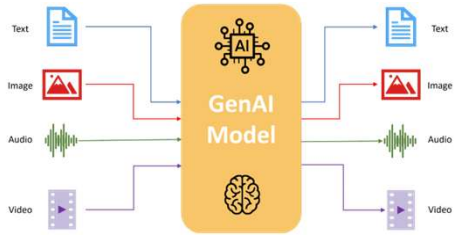


GENERATIVE VS DISCRIMINATIVE MODELS



HOW GENERATIVE MODELS WORKS?

GENAI MODELS : AN OVERVIEW

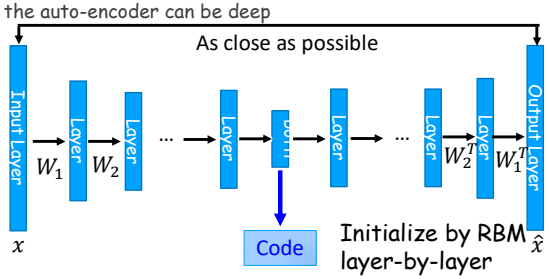


- Encoder – Decoder Models
- Generative Adverserial Networks
- Transformers
- Vision Transformers
- Diffusion Models

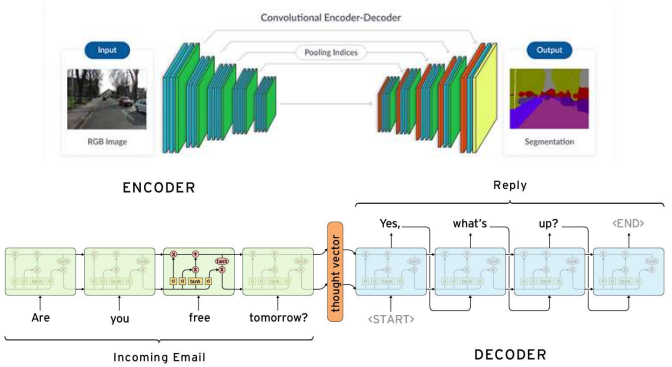
DEEP AUTO-ENCODER

Symmetric is not necessary.

Of course, the auto-encoder can be deep



ENCODER-DECODER



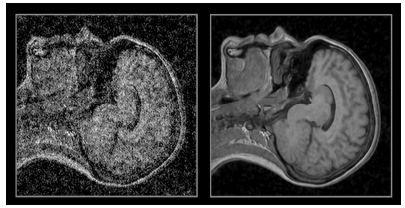
ENCODER-DECODER(VISION)

(a) input image

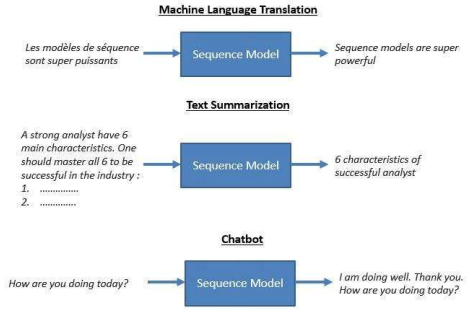
(b) semantic segmentation

(c) instance segmentation

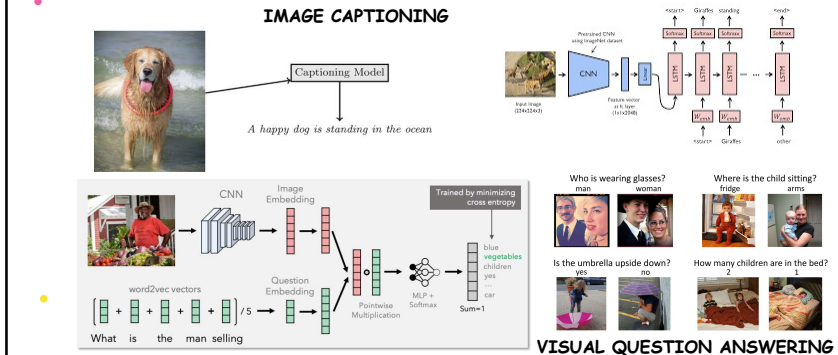
(d) panoptic segmentation



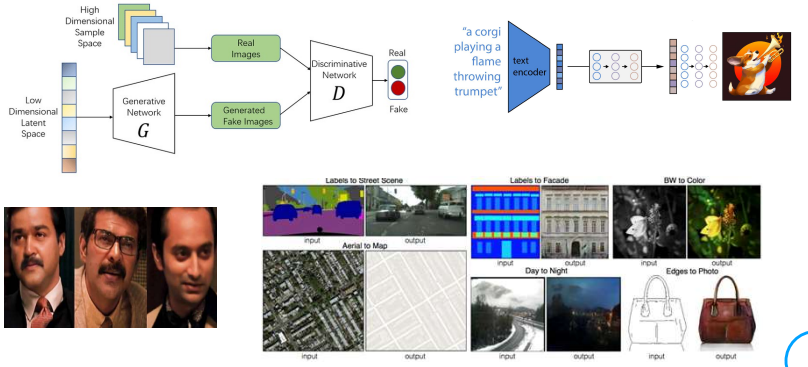
ENCODER-DECODER(TEXT)



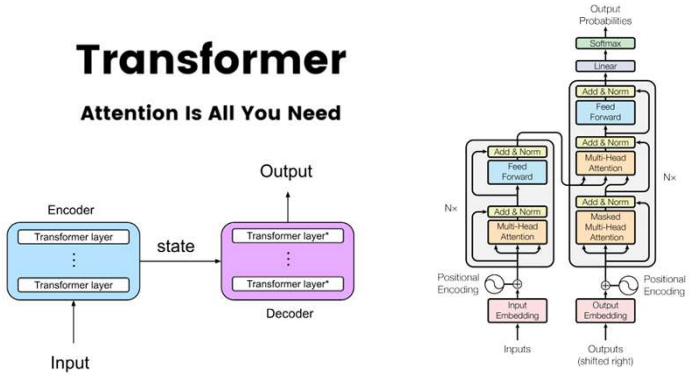
ENCODER-DECODER(TEXT-VISION)



GENERATIVE ADVESERIAL NETWORKS(GANS)



THE SECRET : TRANSFORMER



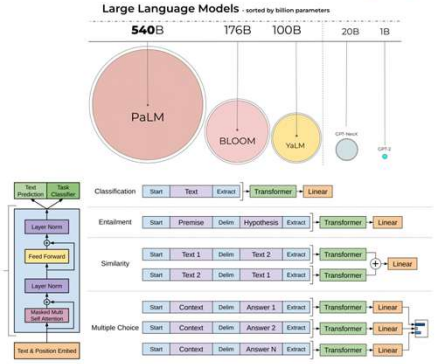
TRANSFORMER TO GPT(AN LLM)

Generative Pretrained Model(GPT)

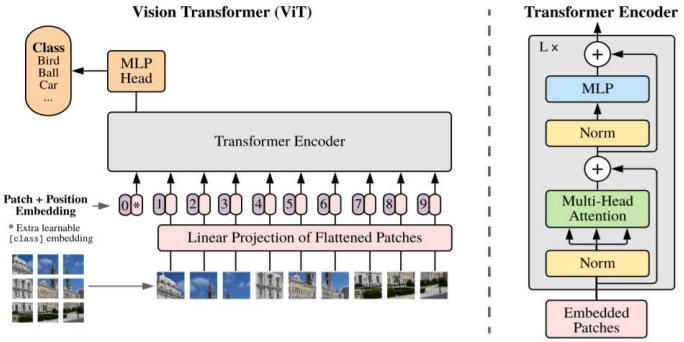
GPT-3 (Generative Pre-trained Transformer 3) is a language model that was created by **OpenAI**.

The 175-billion parameter deep learning model capable of producing human-like text and was trained on large text datasets with hundreds of billions of words.

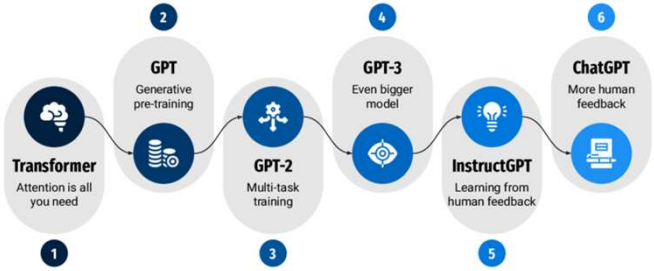
GPT uses an unmodified Transformer decoder, except that it lacks the encoder attention part.



THE SECRET : TRANSFORMER



Evolution from Transformer architecture to ChatGPT

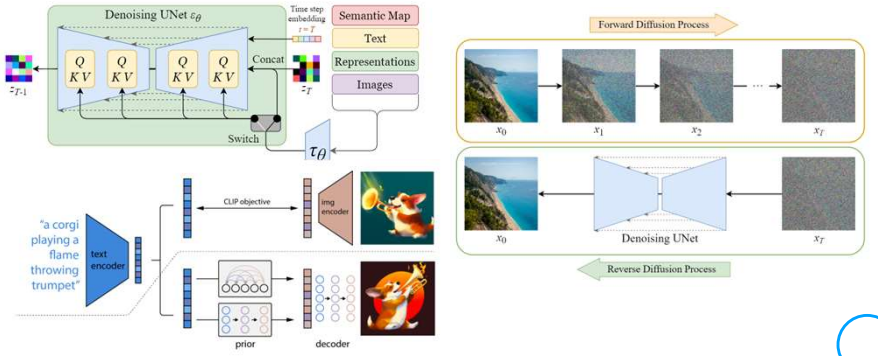


THE SECRET : TRANSFORMER

Vision transformers have extensive applications in popular image recognition tasks such as

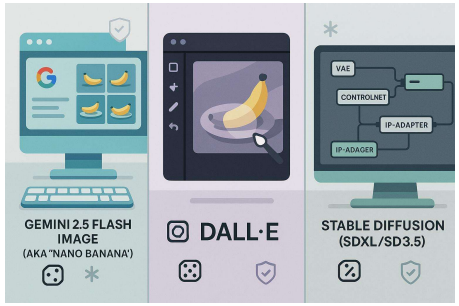
- Object detection
- Segmentation
- Image classification
- Action recognition

DIFFUSION

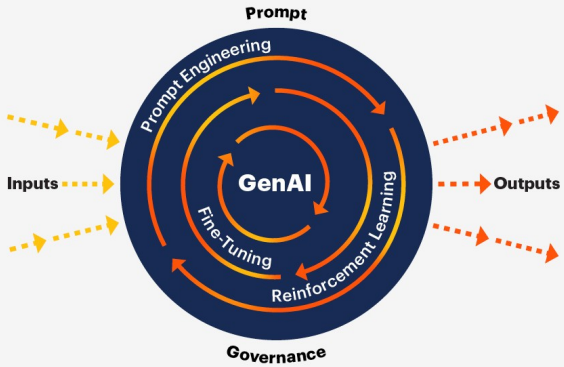


DIFFUSION

- DDPM
- DDIM
- Stable Diffusion
- Latent Diffusion Models (LDM)
- Score-Based Models
- ControlNet



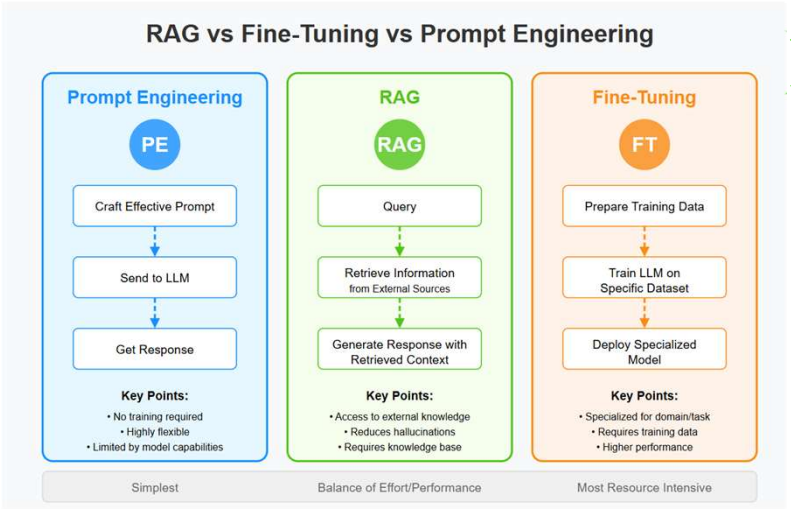
GenAI Framework



Source: Gartner

Gartner

RAG vs Fine-Tuning vs Prompt Engineering



POPULAR GEN AI PRODUCTS

Model	Organization	Use cases
GPT-4	OpenAI	Content generation, code generation, creative writing, question answering
ChatGPT	OpenAI	Dialogue generation, chatbot development
DALL-E 2	OpenAI	Product design, concept art, image editing
Whisper	OpenAI	Transcription, translation, accessibility
Megatron-Turing NLG	Meta	Content generation, code generation, creative writing, question answering
Jurassic-1 Jumbo	Meta	Factual language modeling, topic summarization, question answering, translation
Flan-T5	Meta	Code generation
Bard	Google AI	Content generation, code generation, creative writing, question answering
PaLM	Google AI	Factual language modeling, topic summarization, question answering, translation
Meena	Google AI	Dialogue generation, chatbot development
Turing NLG	Microsoft	Content generation, code generation, creative writing, question answering
CodeGPT	Microsoft	Code generation
DialoGPT	Microsoft	Dialogue generation, chatbot development
Copilot	GitHub	Code generation suggestions
Midjourney	Midjourney	Art and design, product design, education
LLaMA	Meta	Content generation, code generation, creative writing, question answering

Thank You!

Your Queries Please!!!