AI and ChatGPT

Rashid Siddiqui

Chat GPT and AI chatbots

↓

Prompt engineering

What is AI and Machine learning ?

↓

Artificial Intelligence

AI is the capability of a computer system to mimic human cognitive fxns such as learning and problem solving

In year 1997 a robot was made by the name Deep Blue which defeated Garry Kasparov

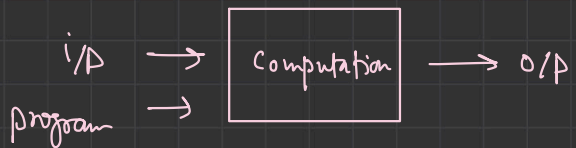
ML

↳ Machine learning → is a subset of AI

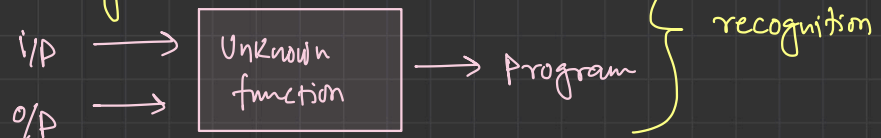
NLP → Natural language processing → To let a computer understand human natural language

ANN → Artificial neural network → To imitate human neural network

Traditional computer programming



Machine learning



AI Chatbots → Rule based chatbots
→ Retrieval based Chatbots
→ Generative Chatbots

The category of AI chatGPT is categorised in is called "Chatbots category"

Simulates human conversations → SIRI
→ Alexa

is generative chatbots category

This is a new category

Techniques used in ChatGPT

Transformers

RLHF

Reinforcement learning & human-feedback

Large datasets are used to train ChatGPT ~ LLM → large language model

Pattern recognition

Next Token Prediction

Masked language modeling

LSTM → long short term memory
is used in predictive texts like in imessages
lacks contextual understanding. Processes the words individually

To solve this issue Google introduced Transformers in 2017 → Google Brain

Processes all input data at once

The novelty of ChatGPT

Open AI launched its first **Generative Pre-training Transformer** model in 2018

↓

named as **GPT-1**

ChatGPT combines "Chat", referring to its chatbot functionality, and "GPT", which stands for "Generative pre-training Transformer", a type of LLM

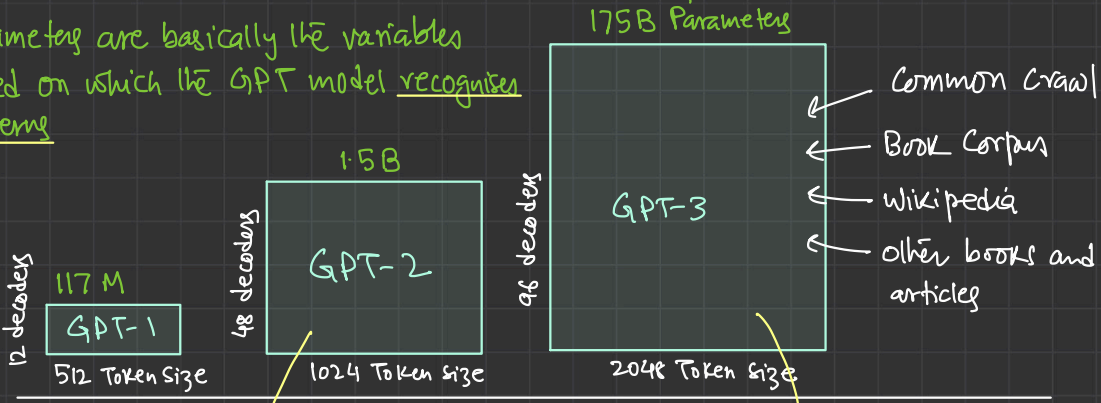
↓

used two datasets

- Common Crawl → millions of webpages
- 11000 books → Book Corpus

Feb 2019 → **GPT 2.0** was launched

Parameters are basically the variables based on which the GPT model recognises patterns



Year 2020, **GPT-3** model was launched

- → 1.5B parameters
- → 40GB text training dataset
- → often fine tuned to perform specific tasks
- → smaller version of the model was released to the public open source

- → 176B parameters
- → 570 GB training dataset comp of books, articles, websites etc.
- → Ability to perform most language tasks w/o additional tuning
- → launched as an API service

Nov 2022- GPT 3.5 is launched

↳ in this version RLHF was used

↓

Reinforcement learning c human-feedback

Uses PPO

Proximal policy optimisation

There are three main steps in RLHF

1. Supervised fine tuning model
2. Reward model
3. Reinforced learning model

Couldn't started rewarding self

Training data for GPT3 and GPT3.5 is same

Just the coding of GPT 3.5 is different

To fine tune 40 contractors were hired

For many inputs, these people generated manual output by typing the corrections

↓

hence fine tuning

Very slow

Time consuming

Costly

manually ranking answers generated by ChatGPT

↓

high score answers

March 2023 — GPT 4 ← Trained data till sep 2021

Number of parameters are not revealed yet

+

training data

↳ assumed to be trillion of parameters

Chat GPT model is evaluated based on three criteria

↓

4.0

Helpfulness

Truthfulness

Harmlessness

Chatbots and prompt engineering

Chat GPT

4.0 is paid

Bing AI

same data set as used by Bing AI

→ You can get 4.0 advantage free in Bing AI

→ Can cutoff sep 2021 limit

Google Bard

inferior to GPT

Now, we need experts to smartly handle prompts for business

Prompt engineering → to initiate a conversation

We used to write programs in computer languages to talk to computers but in AI, we can use natural language to interact to computers

Tokenizer tool → will tell you tokens and characters count

Tokens → "Smallest unit of information understood by AI"

Chat GPT doesn't recognise words, rather it recognises **tokens**

Tokens are the individual units of information through which chat GPT understands information

It is important to know this (Tokens) because the limit set by OpenAI for chat GPT is based on tokens.

- In chat GPT 3.5 there is a **limit of 4096 tokens** for one prompt
  - including both question and answer
- In Chat GPT 4.0 the limit is 32768 tokens
  - ↳ ~ 24000 english words

Effective Prompt engineering

1. Make CG pretend to be an expert
2. Define objective very specific
3. Ask output format
5. Prompt by example
6. Ask CG to ask you questions
7. Refine your answer and counter question.

AI-900

Azure
↓
200+ services

AI and ML
- "General AI"
  - ↓ Strong AI
- Narrow AI ← also called weak AI
  - ↳ Focus on specific task

Self driving cars
Spam filters
Email classification
Fraud detection

ML is a category of Narrow AI

Machine learning AI
↓
Computer vision
↓
- → Analyse & describe
- → Read text in imagery
- → Read handwriting
- → Recog celeb & landmarks

Reinforcement learning
↳ program become its own teacher

Machine learning vs Traditional programming

Machine learning
↓
learning from examples

Traditional programming
↓
if this do that
learning from Rules

Steps.
1. Give millions of examples
2. Create a model
3. Use the model to make predictions

↑ need a lot of data
skilled person
Complexity

Three approaches to building AI solutions in "Azure"

1. Use pre-trained models – Azure Cognitive services
   ↓
   text, image, audio video
2. Build simple models – without needing data scientists
   – limited/no-code experience
   – Custom vision
   – Azure Machine learning ← Also called Automated m/c learning
3. Build complex models – using data scientists
   – Using Azure m/c learning

AI workload types

→ Content moderation and Personalisation

→ Computer vision workloads

→ Natural language processing workloads

→ Knowledge mining workloads

→ Document intelligence workloads

→ Generative AI workload

Exploring pre-trained models

↓

Cognitive Services ← bring AI within reach of every developer

renamed as "Azure AI services"

– AI w/o building custom models

– No need of m/c learning expertise

– Exposed as API