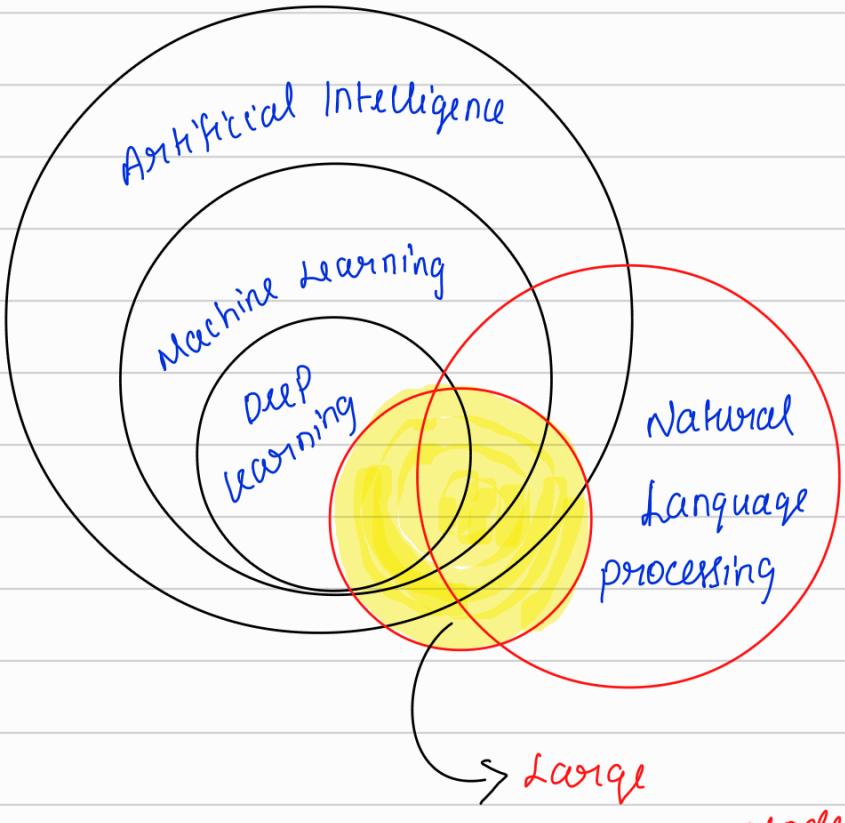
Large Language Models (LLMs) concept

Introduction to LLM

AI performed well in data driven tasks.
(Sentiment analysis and fraud detection)

Still, lacked human-like interaction.

So, Large language Model came.



Definition

* Large (Training data & compute power)
* Language (Human-like text)
* Model (Learn complex patterns using text data)

Applications

- Sentiment analysis
- Identifying themes
- Translating text / speech
- Generating code.
- Next word prediction.

Real word application

Business Opportunities

Business constantly seek new and innovative ways to improve their products and services using LLM.

Benefits.

- Automate tasks
- Improve efficiency
- Create revenue stream
- Enable new capabilities.

Transforming finance industry

Investment outlook | Annual report | News articles | Social media posts.

↓

Processing a unstructured data (data that lacks definion and is presented free form)

LLM can analyse such data.

challenges in Healthcare

Analyzing Health records is important for giving personalized recommendations to provide quality healthcare.

Analyze patient record to offer personalized recommendation.

→ Must adhere to privacy laws (data)

Medical record

Health checkup → LLM → personalized treatment recommendation.

Imaging report

Education

→ personalized coaching & feedback.
→ Interactive learning experience

→ AI-powered tutor
(Ask questions)
(Receive Guidance)
(Discuss Ideas)

→ personalizing education (text generation)

Multimodal
↳ Many types of processing or generation
(text & audio & image)

→ visual question answering is one such Multimodal application (Answers to questions about visual content)

→ object identification & relationship
→ Scene description
→ Responds with additional images

Challenges of language Modelling

1. sequence matters
2. context modeling.
3. Long range dependency (Recognize and connect distant words in a sentences)
↳ challenging for traditional language models.

Single task learning (traditional model)

- Time and resources expensive
- less flexible compared to modern LLMs.

Multitask learning. (Modern LLM)

- Improved performance on each individual tasks.
- might impact accuracy and efficiency.
- less training data needed because data is shared.

Novelty of LLMs (unique capabilities)

→ NLP techniques are the foundations of LLMs.

→ Linguistic subtleties.
(irony, humor, intent, sarcasm...)

overcome data's unstructured nature.

How do LLMs understand:

→ Trained on vast amount of data.
→ Largeness of LLMs (parameter)
→ parameter represent the patterns and rules
→ more parameters → complex patterns
→ Generates sophisticated and accurate responses.

Emergence of new capabilities

* Emergent abilities
  (only present in large scale model)
  → Music
  → Poetry
  → code generation
  → Medical diagnosis & treatment plans

* Scale,
  Volume of training data,
  Number of Model parameter,

Building Block of LLMs

→ Text pre-processing
→ Text representation
→ pre-training
→ Fine-tuning
→ Advanced fine-tuning

Generalized Overview of NLP

① Text Pre-processing

* tokenization
* Lemmatization
* Stop word removal.

can be done in different order.

Split text into individual words or token

Stop words do not add meaning & eliminated

* Group slightly different words with similar meanings
* Reduces words to their base form.
* Mapped to root word.

(2.) Text Representation.

(Text data into numerical form)

* Bag of words
* word embeddings.

Text into a matrix of word counts.

'0' represents the absence of a word.

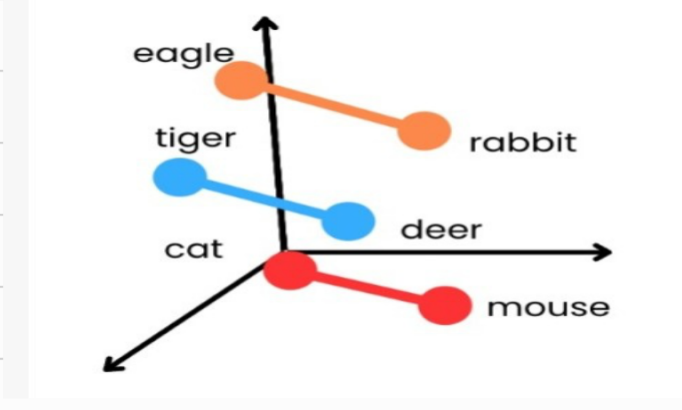
Limitation

→ does not capture the order / context

→ does not capture the semantics between the words.

capture the semantic meanings as number.

predator - prey relation



start with text pre-processing

Machine Readable form

convert pre processed text into numerical form.

Text data → Text Pre-Processing → Text Representation → Numerical data

③ pre-training.

Not everyone needs to train an LLM from scratch

- **Pre-training**

School education

- **Fine-tuning**

University specialization

Largeness challenge

* fine-tuning can help
* powerful computers

computing power.

→ memory
→ processing power
→ infrastructure
→ Expensive.
→ LLM, 10K of CPU & GPU

* Efficient model training methods
* Large amounts of training data.

Efficient model training.

→ Training time is huge
→ may take weeks or even months
→ Efficient model training = faster training time.
→ 355 years of processing time on a single GPU.

Data availability

→ Need of high-quality of data
→ to learn the complexities & subtleties of language.
→ few 100's GB of text data.
→ Massive amount of data.

Overcoming the challenges

- **Fine-tuning**
  - Addresses some of these challenges
  - Adapts a pre-trained model

- **Pre-trained model**
  - Learned from general-purpose datasets
  - Not optimized for specific-tasks
  - Can be fine-tuned for a specific problem

Fine - tuning.	pre-training.
* compute (1-2 CPU & GPU)	* compute (1000's of CPU & GPU)
* Training time (Hours to day)	* Training time (weeks to Month)
* Data (~1 GBs)	* Data (~100 of GBs)

Learning techniques

fine-tuning, training a pre-trained model for a specific task.

Transfer learning
↳ learn from one task and transfer to related task.

N-shot learning
↳ zero shot, few shot and Multishot learning.

Building Blocks → Data preparation workflow
→ Fine tuning
→ N-Shot Learning techniques

Generative pre-training

* Trained using generative pre-training
  → input data of next token
  → Trained to predict the token with the dataset.

* Types:
  → Next word prediction
  → Masked Language modeling.

Next word prediction

* Supervised Learning technique
  (Model trained on input-output pairs)

* Predicts next word and generates coherent text and captures the dependencies between words.

* Training data (pairs of input & output examples)

* More examples = better prediction.

Training data for next word prediction

Input	Output
The quick brown	fox
The quick brown fox	jumps
The quick brown fox jumps	over
The quick brown fox jumps over	the
The quick brown fox jumps over the	lazy
The quick brown fox jumps over the lazy	dog
The quick brown fox jumps over the lazy dog.	

Masked Language Modelling

→ Hides a selective word

→ Trained model predicts the masked word.

Original Text: "The quick brown fox jumps over the lazy dog."

Masked Text: "The quick [MASK] fox jumps over the lazy dog."

→ Objective, predict the missing data

→ Based on learnings from training data.

Transformers

* Attention is all you need.
  (Revolutionized language modelling)

* Transformer architecture.
  → Relationship between words
  → Components, pre-processing, positional encoding, Encoders & decoders.

input → Text pre-processing & representation → positional encoding → Encoders → Decoders → output.

Positional encoding

→ Information on the position of each word.
→ understand distant words.

Encoders

→ Attention Mechanism, directs attention to specific words and relationship
→ Neural network, process specific features.

Decoders

→ includes attention and neural network.
→ Generates the output

Transformers & long-range dependencies

→ Initial challenges, long range dependency.
→ Attention, focus on different parts of the input.

Processes Multiple parts simultaneously.

→ Limitation of traditional language model.
(Sequential – one word at a time)
→ Transformer,
* process Multiple parts simultaneously.
* faster processing.

Attention Mechanism

→ understand complex structures
→ focus on important words

Self attention

→ weights the importance of each word.
→ captures long range dependencies

Multi-head attention

→ Next level of self attention
→ Splits inputs into multiple heads with each head focussing on different aspects.

Attention in party

Self-attention

- Focus on each person's words
- Evaluate and compare their relevance
- Weigh each speaker's input
- Combines for a comprehensive understanding

Multi-head attention

- Split attention into "**multiple**" channels
- Focus on different aspects of conversation
- Speaker's emotions, primary topic, and related side-topics
- Process each aspect and merge

Multi-head attention advantages

example

- "The boy went to the store to buy some groceries, and he found a discount on his favorite cereal."

- **Attention**: "boy," "store," "groceries," and "discount"
- **Self-attention**: "boy" and "he" -> same person
- **Multi-head attention**: multiple channels
  - Character ("boy")
  - Action ("went to the store," "found a discount")
  - Things involved ("groceries," "cereal")

Multi head attention is like having multiple self attention mechanisms working simultaneously

Advanced fine-tuning

Reinforcement Learning through Human feedback

* pre-training
* Fine-tuning
* Reinforcement Learning through Human feedback (RLHF)

pre-training. →
- → large amounts of text data
- → Next word prediction
- → Masked Language Modelling.

Fine-tuning →
- → N-Shot training
- → small labeled dataset for related task.

Why RLHF?

→ General purpose training data lacks quality (Noise, Errors, inconsistencies & reduced accuracy)

Pre training	Fine tuning	RLHF !
↓	↓	↓
learns underlying language patterns	Quality labled data improves performance.	Human Feedback.
Doesn't capture context complexities		

Simplifying RLHF

→ Model output reviewed by human
→ updates model based on the feedback.

Step 1:
→ Receives a prompt
→ Generates Multiple responses.

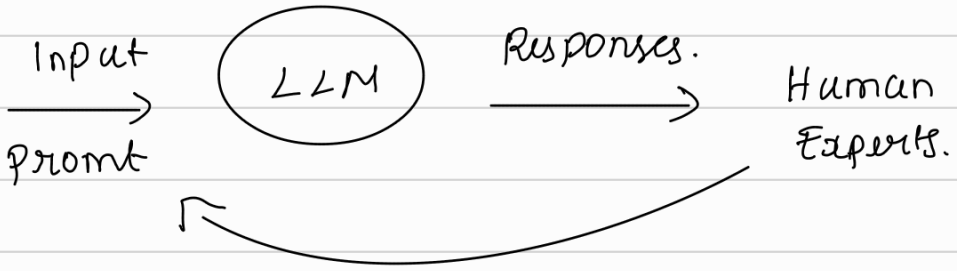
Step 2:
→ Human expert checks these responses
→ Ranks the responses based on quality.
  * Accuracy
  * Relevancies
  * Coherence.

} Enter human Experts.

Step 3:
→ Learn from expert's ranking
→ To align its responses in future with their preferences.

And it goes on!
* continues to generate responses
* Receives expert's ranking
* Adjusts the learning.

[RLHF techniques to enhance fine-tuning through human feedback]



Data concerns and consideration

Data considerations

* Data volume & compute power → LLM need a lot of data, similar to a child learning to talk (570 GB, ~1.3M books)
* Data quality
* Labelling
* Bias
* privacy.

Accurate data = better learning = improved response quality = increased trust.

→ Extensive computing power, think of the energy consumption

→ can cost millions of dollars!

Labelled data

→ correct data label. (accurate learning, generalize patterns, accurate response)

→ Labor-intensive (assigning correct label to each article)

→ Incorrect labels impact model performances.

→ Addresses errors: identify → analyse → iterate.

Data bias

→ influenced by social stereotypes

→ Lack of diversity in training data

→ Descrimination & unfair outcomes.

→ spot and deal with the biased data
* Evaluate data imbalance
* promote diversity
* Bias mitigation techniques : more diverse examples.

Data privacy

→ compliance with data protection & privacy regulation
→ sensitive / personally identifiable information (PII)
→ Get permission
→ privacy is a concern
* Training on data without permission can lead to a breach
* Legal, financial and reputation harm

Ethical concern

* Transparency risk. → challenging to understand the Output
* Accountability risk
* Information hazards.
  - Disseminating harmful information
    → Harmful content generation,
    → Misinformation spread,
    → Malicious use,
    → Toxicity
  - Responsibility of LLM's actions.

Environmental concerns

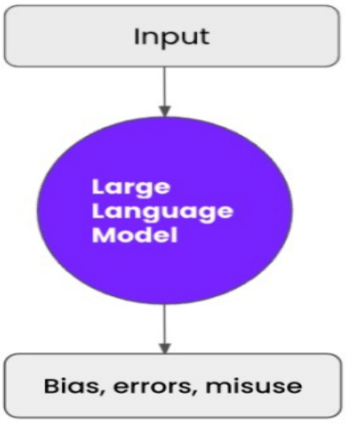
→ Ecological footprints of LLMs
→ substantial energy resources to train
→ Impact through carbon emissions.

Cooling requires electricity too!

- Produce considerable heat that needs cooling
- Imagine thousands of laptops overheating
  - Require complex cooling systems
  - Adds to environmental impact
- Balance the cost and benefits
  - Use renewable energy
  - Energy-efficient tech

Transparency risk

- Challenging to understand the output
- Difficult to identify issues
  - Bias
  - Errors
  - Misuse
- Black box
- Example: reasoning behind predicting disease outcomes



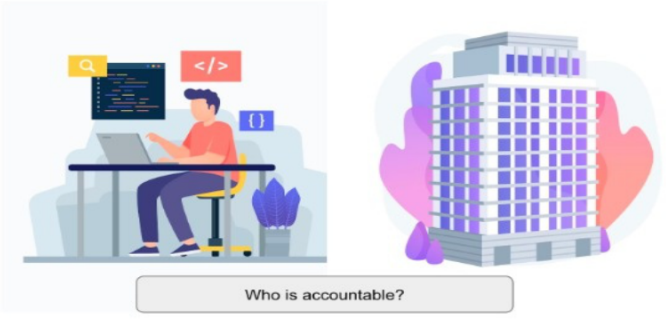
↳ LLM must explain for making informed treatment decisions

Accountability risk

- Responsibility of LLMs' actions

- Who is responsible?
  - Incorrect and harmful advice
  - Model developer or the company?

- Game without rules
  - No transparency
  - No accountability



where are LLMs headings ?

1. Model explainability.
   - → How do they arrive at their outputs.
   - → Builds trust and transparency
   - → Identify and correct the biases or errors.

2. Efficiency.
   - → computational efficiency.
     (High quality output with less compute)
   - → Faster & efficient
     (Model compression & optimization)
   - → Benefits.
     (Better storage, lower energy use)
   - → Accessibility & Sustainability
     (promotes green AI,
     Reduces operating costs)

3. Unsupervised Bias Handling.
   - → Biased data (discrimination)

→ unsupervised bias handling.

* Bias detection and mitigation techniques automatically
* No need of explicity human labelled data.
* identifies & reduces by analysing patterns

→ Challenges

* Subtles, difficult to detect
* Might introduce new biases.

4.

Enhanced creativity

- Creativity in text-based and visual art forms
- **Artistic content**: learned patterns, not emotional understanding
- Lack human-like comprehension of art or emotions
- Demonstrate human-like emotional behavior
- **Future**: emotion inference