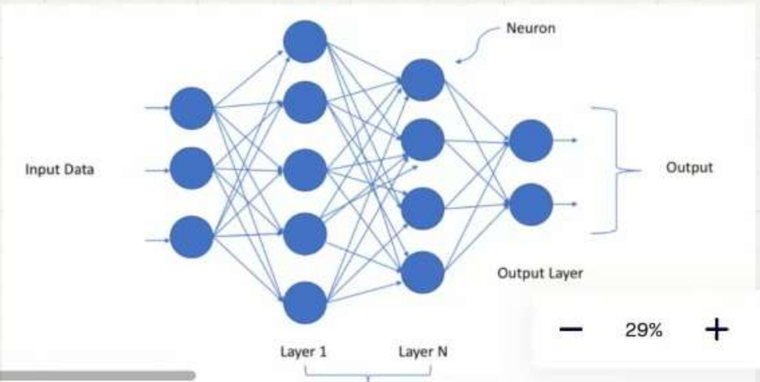
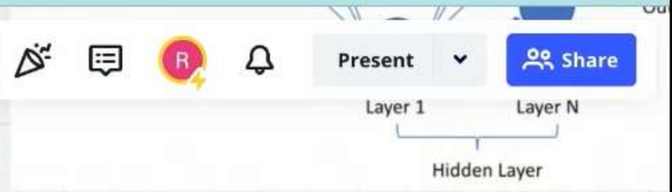
Lecture 2: Intro to LLMs

① What is a Large Language Model (LLM)

Neural networks designed to understand, generate and respond to human like text.

⇒ Deep Neural Networks trained on massive amounts of text data





② LARGE LANGUAGE MODELS (LLMs)

Models have billions of parameters!

These models do a wide range of NLP tasks: question answering, translation, sentiment analysis and much more!

③ LLMs vs. Earlier NLP models

much more!

③ LLMs vs Earlier NLP models

can do wide range of NLP tasks

designed for specific tasks like language translation etc

"Earlier language models could not write an email from custom instructions, a task

range of NLP tasks

tasks like language translation etc

"Earlier language models could not write an email from custom instructions, a task that is trivial for modern LLMs"

④ What makes LLMs so good? Secret sauce?

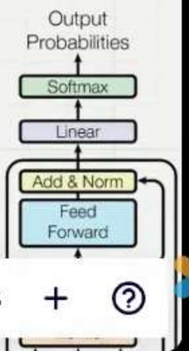
miro | Convolutional Neural Networks

Upgrade | Present | Share

"Earlier language models could not write an email from custom instructions, a task that is trivial for modern LLMs"

④ What makes LLMs so good? Secret sauce?

Transformer Architecture (What does it mean)



33%

miro.com/app/board/uXjVKENK1Qc=/

Finish update

miro | Convolutional Neural Networks | Upgrade | Present | Share

Transformer Architecture

(What does it meon)

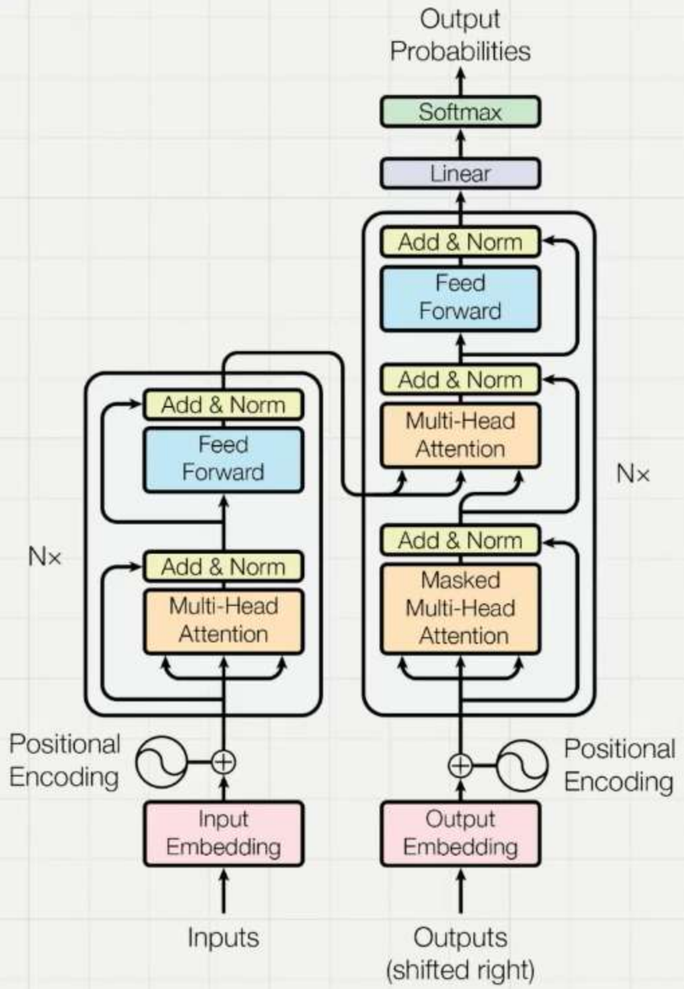
Nx

Positi
Encoc

(This is the only Transformer

44%

Transformer



(It actually something this)

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer,

miro.com/app/board/uXjVKENK1Qc=/

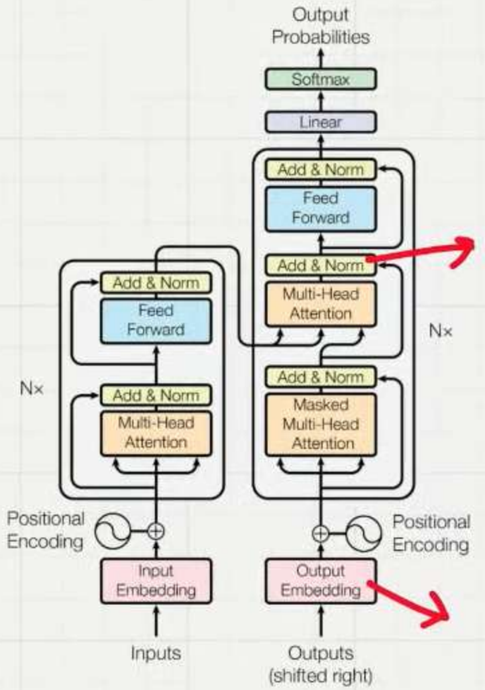
Finish update

miro | Convolutional Neural Networks | Upgrade | Present | Share

Architecture

(What does it mean)

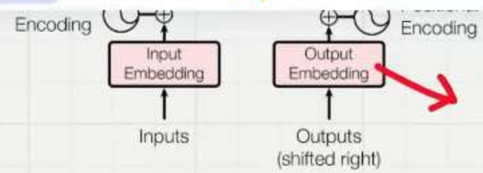
(This is the only Transformer I know)



(It som

32%

(This is the only Transformer I know)

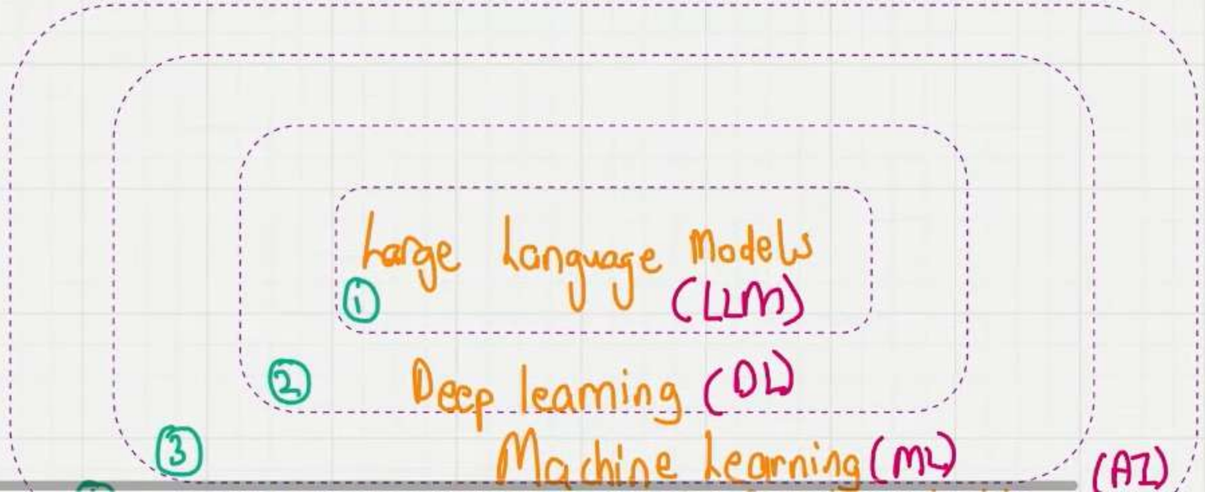


Don't worry. We will learn all about this secret sauce: Transformers

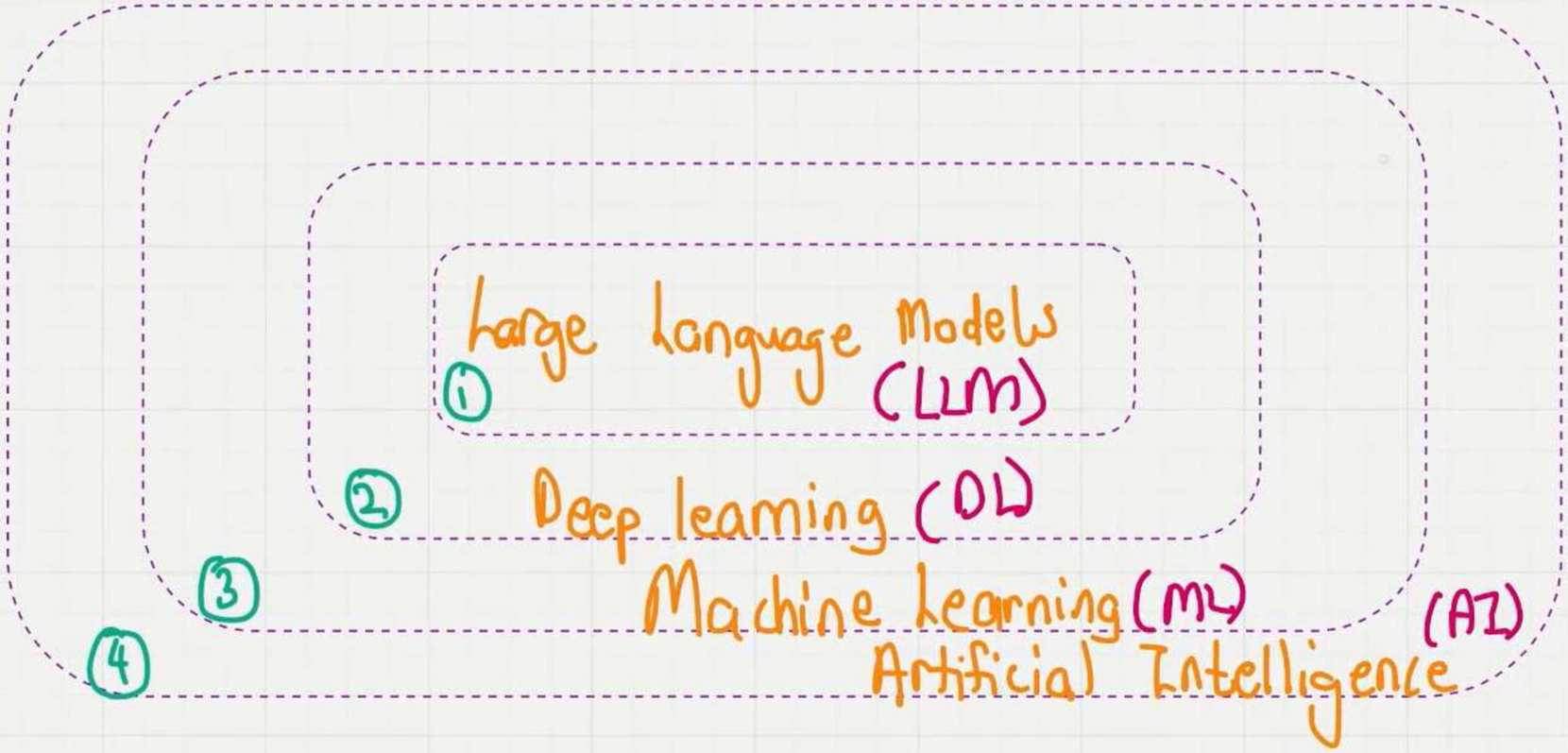
(5) LLM vs Gen AI vs Deep learning vs Machine

Don't worry. We will learn all about this secret sauce: Transformers

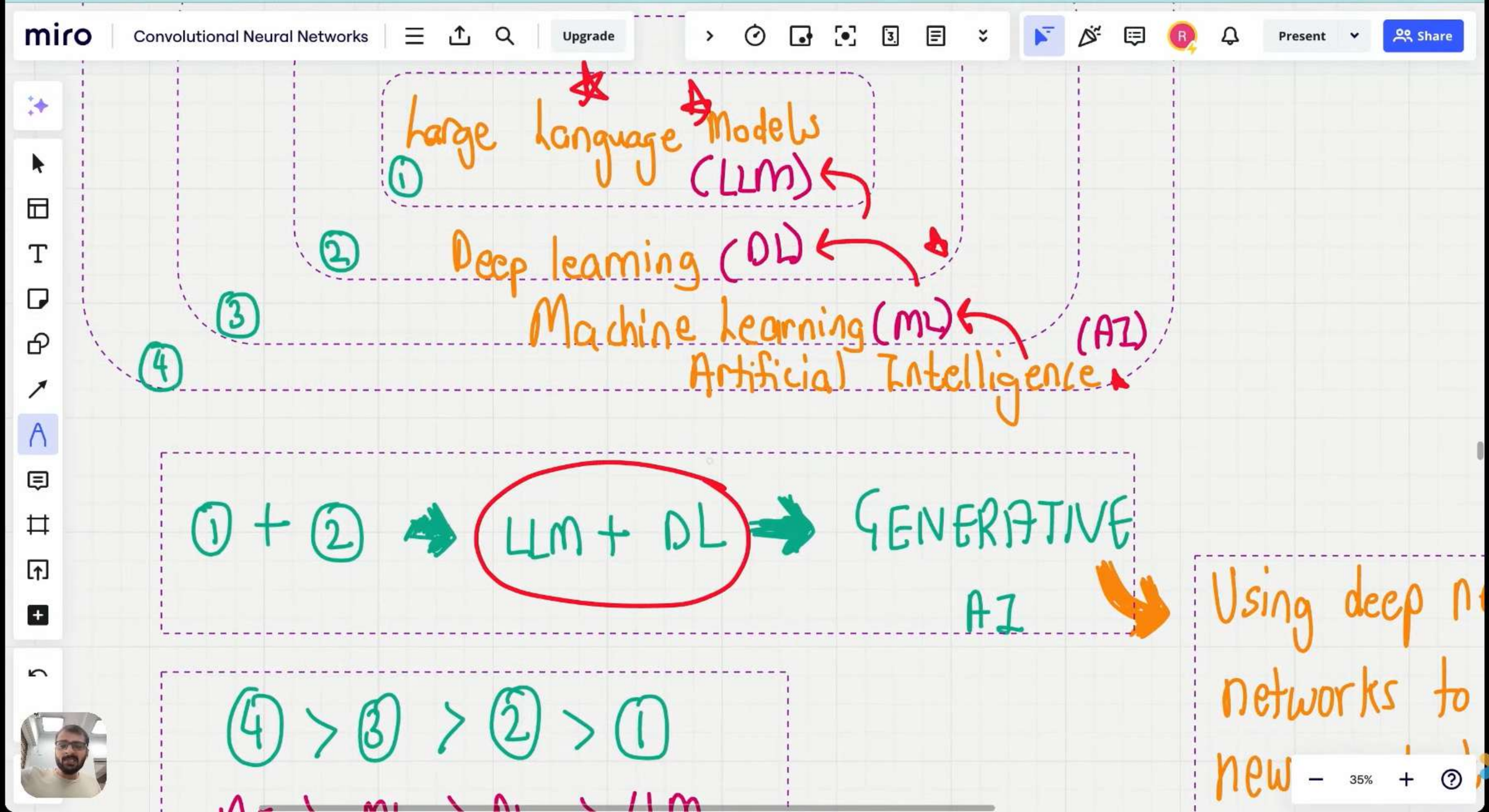
(5) LLM vs Gen AI vs Deep learning vs Machine learning



LLM vs Gen AI vs Deep learning vs Machine learning



1 + 2 = LLM + DL = GENERATIVE


miro
Convolutional Neural Networks
Upgrade
Present
Share
Large language Models (LLM)
Deep learning (DL)
Machine learning (ML)
Artificial Intelligence (AI)
1 + 2
LLM + DL
GENERATIVE AI
4 > 3 > 2 > 1
Using deep n
networks to
new
35%

miro.com/app/board/uXjVKENK1Qc=/

miro | Convolutional Neural Networks | Upgrade | Present | Share

Artificial Intelligence (AI)

LLM + DL → GENERATIVE AI

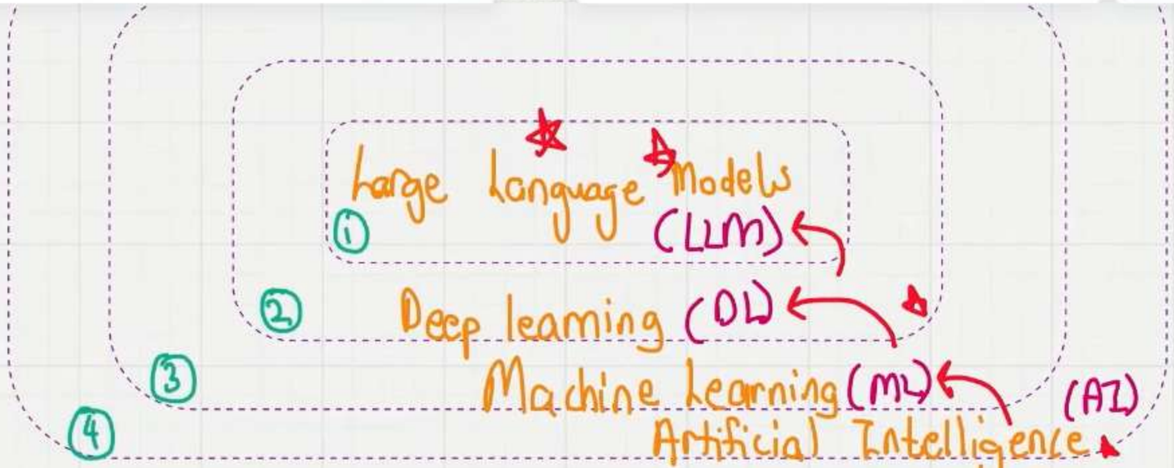
Using deep neural networks to create new content such as text, images, various forms of media.

② > ①

DL > LLM

t a specific application

35%



① + ② → LLM + DL → GENERATIVE AI

Using deep neural networks to create new content such as text, images, various forms of medi

④ > ③ > ② > ①

AI > ML > DL > LLM

LLMs represent a specific application of deep learning techniques leveraging

networks to cre
new content su
images, variou

④ > ③ > ② > ①

AI > ML > DL > LLM

LLMs represent a specific application of deep learning techniques, leveraging their ability to process and generate human like text.

⑤ Applications of LLMs

their ability to process and generate human like text ★

6 Applications of LLMs

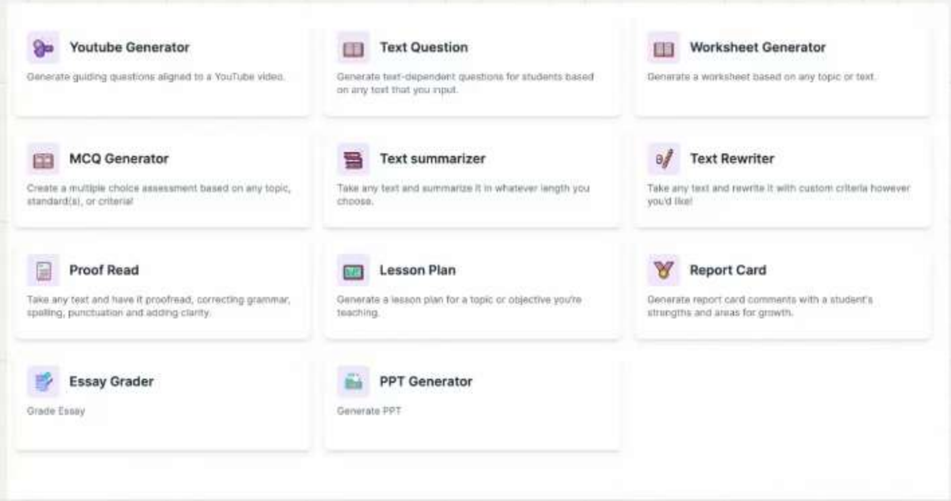
- Chatbots/ Virtual Assistants
- Machine translation
- novel text generation
- Sentiment analysis
- Content creation

human like text

Applications of LLMs

6

1. Content creation
2. Chatbots/ Virtual Assistants
3. Machine translation
4. novel text generation
5. Sentiment analysis



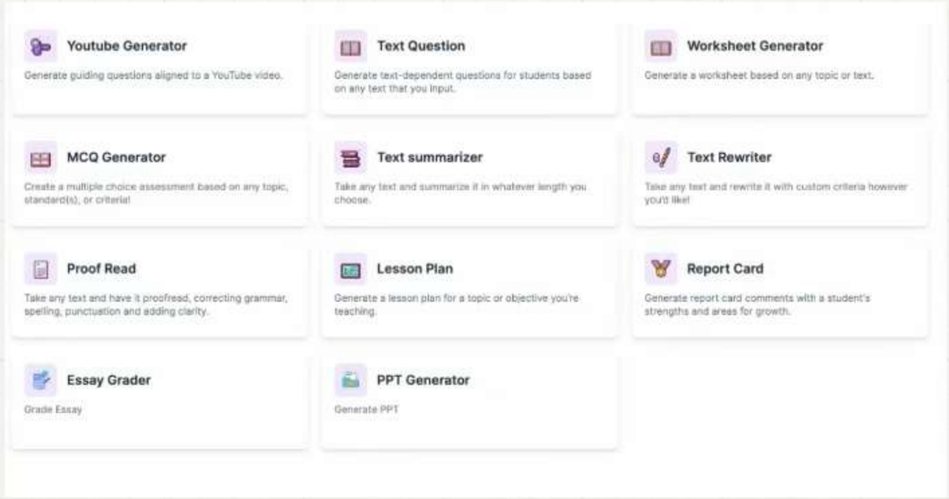
miro.com/app/board/uXjVKENK1Qc=/

Finish update

miro | Convolutional Neural Networks | Upgrade | Present | Share

Assistants

Translation



"The sky is the limit when it comes to LLM applications"

26%

Lecture3 : Stages of building LLMs

Creating an LLM = Pretraining + Finetuning

Training on a large, diverse dataset

Refinement b training on dataset

Lecture3 : Stages of building LLMs

Creating an LLM = Pretraining + Finetuning

Training on a large, diverse dataset

Refinement by training on narrowe dataset, specific to particular task or

	Proportion in ining data
	%
	%

Pretraining + Finetuning schematic:

Pretrained L

Dataset name	Dataset description	Number of tokens	Proportion in training data
CommonCrawl (filtered)	Web crawl data	410 billion	60%
WebText2	Web crawl data	19 billion	22%
Books1	Internet-based book corpus	12 billion	8%
Books2	Internet-based book corpus	55 billion	8%
Wikipedia	High-quality text	3 billion	3%

Creating

Total pre-training cost for GPT-3

Pretraining +

miro.com/app/board/uXjVKENK1Qc=/

miro | Convolutional Neural Networks | Upgrade | Present | Share

The lion is in the forest

Dataset name	Dataset description	Number of tokens	Proportion in training data
CommonCrawl (filtered)	Web crawl data	410 billion	60%
WebText2	Web crawl data	19 billion	22%
Books1	Internet-based book corpus	12 billion	8%
Books2	Internet-based book corpus	55 billion	8%
Wikipedia	High-quality text	3 billion	3%

Creating

...tal pre-training cost for GPT-3

Pretraining

27%

upervised learning?

ks

: Dataset examples

Sentiment Analysis

Winograd Schema Resolution

Linguistic Acceptability

Question Answering

Pre-trained

Random Init

We also noticed we can use the underlying language model to begin to perform tasks without ever training on them. For example, performance on tasks like picking the right answer to a multiple choice question steadily increases as the underlying language model improves. While the absolute performance of these methods is still often quite low compared to the supervised state-of-the-art (for question answering it still outperformed by a simple sliding-window baseline) it is encouraging that this behavior is robust across a broad set of tasks. Randomly initialized networks containing no information about the task and the world perform no-better than random using these heuristics. This provides some insight into why generative pre-training can improve performance on downstream tasks.

We also noticed we can use the underlying language model to begin to perform tasks without ever training on them. For example, performance on tasks like picking the right answer to a multiple choice question steadily increases as the underlying language model improves. While the absolute performance of these methods is still often quite low compared to the supervised state-of-the-art (for question answering it still outperformed by a simple sliding-window baseline) it is encouraging that this behavior is robust across a broad set of tasks. Randomly initialized networks containing no information about the task and the world perform no-better than random using these heuristics. This provides some insight into why generative pre-training can improve performance on downstream tasks.

We can also use the existing language functionality in the model to perform sentiment analysis. For the Stanford Sentiment Treebank dataset, which consists of sentences from positive and negative movie reviews, we can use the language model to guess whether a review is positive or negative by inputting the word “very” after the sentence and seeing whether the model predicts the word “positive” or “negative” as more likely. This approach

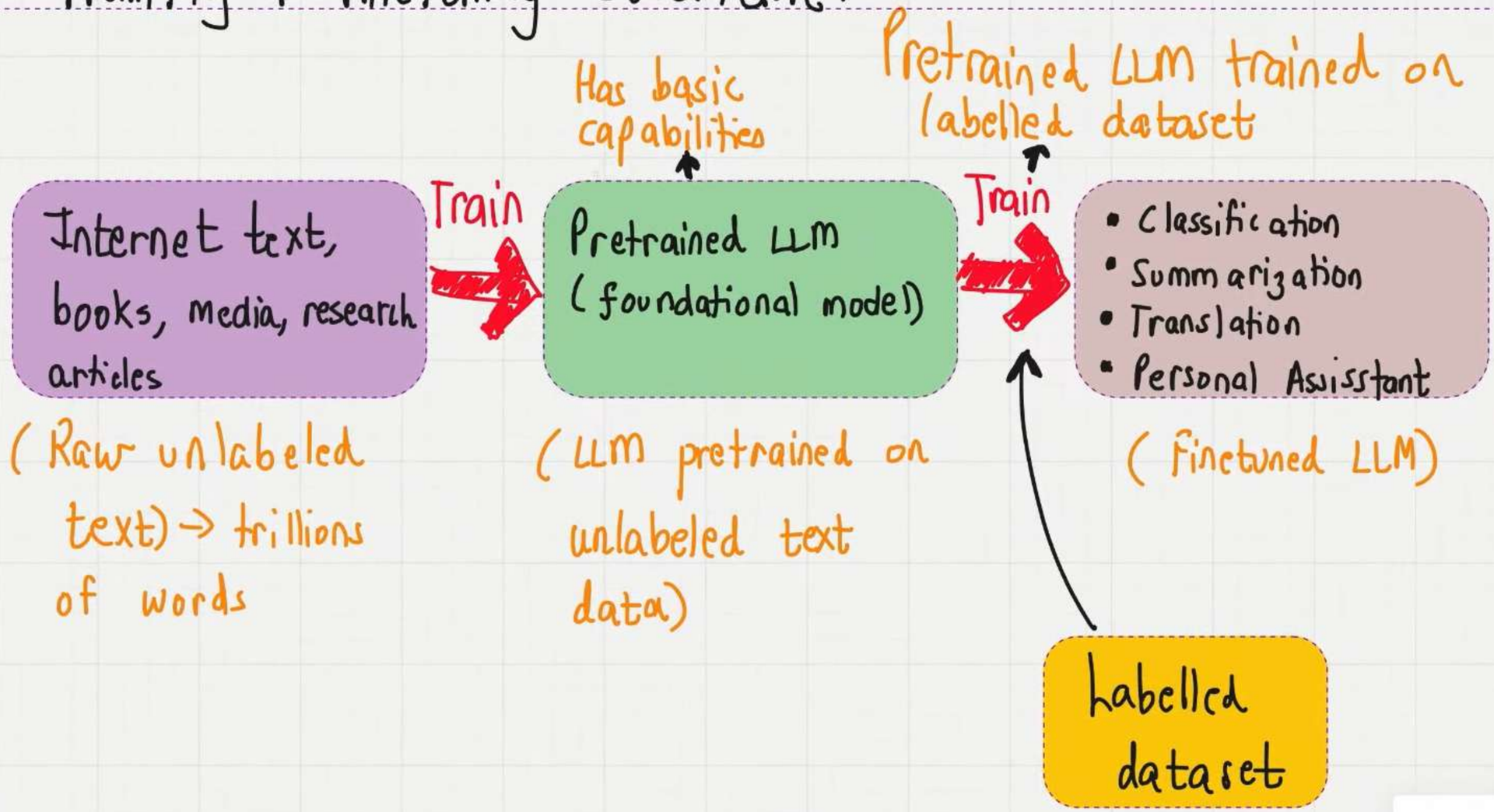
= Pretraining + Finetuning

ining on a

ge, diverse dataset

Refinement by training on narrower dataset, specific to a particular task or domain

Pretraining + finetuning schematic:



dataset

Steps for building a LLM:

① Train on a large corpus of text data (raw text)

Raw text = regular text without any labeling information.

② First training stage of LLM is also called pretraining.

↓ creating an initial pretrained LLM

Raw text = regular text without any labeling information.

② First training stage of LLM is also called pretraining.

↓ creating an initial pretrained LLM (base/foundational model)

Pretr... basic pred...

Example! GPT-3 model is a pretrained model which is capable of text completion.

③ After obtaining the pretrained LLM, we can

(base/foundational model)

Example! GPT-3 model is a pretrained model which is capable of text completion.

basic

pred

(3) After obtaining the pretrained LLM, we can further train LLM on labelled data.

(4) There are 2 popular categories of finetuning

instruction finetuning

finetuning for

(base/foundational model)

Example! GPT-3 model is a pretrained model which is capable of text completion.

basic

pred

(3) After obtaining the pretrained LLM, we can further train LLM on labelled data. → finetuning

(4) There are 2 popular categories of finetuning

instruction finetuning

finetuning for

is capable of text completion.

(3) After obtaining the pretrained LLM, we can further train LLM on labelled data. → finetuning

(4) There are 2 popular categories of finetuning

- instruction finetuning
  - labeled dataset consists of
- finetuning for classification tasks

further train LLM on labelled data. → finetuning

④ There are 2 popular categories of finetuning

instruction finetuning

labeled dataset consists of instruction- answer pairs.

eg: text translation,

finetuning for classification tasks

labeled dataset consists of text & associated labels

instruction finetuning

Labeled dataset consists of instruction- answer pairs.
eg: text translation,
airline customer support

finetuning for classification tasks

Labeled dataset consists of text & associated labels.
eg: emails → spam vs

miro.com/app/board/uXjVKENK1Qc=/

miro | Convolutional Neural Networks | Upgrade | Present | Share

There are 2 types of categories of finetuning

instruction finetuning

eled dataset consists of
ruction- answer pairs:
ext translation,
airline customer support

finetuning for classification tasks

labeled dataset consists of text & associated labels.
eg: emails → spam vs no-spam.

30%