

Conv x 033b x 1_laD x d415 x Chat x 1706 x rule-t x Build x Heart x CNN x Lab P x vizua x + miro.com/app/board/uXjVKENK1Qc=/ Finish update :

miro Convolutional Neural Networks Upgrade DM Present Share

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

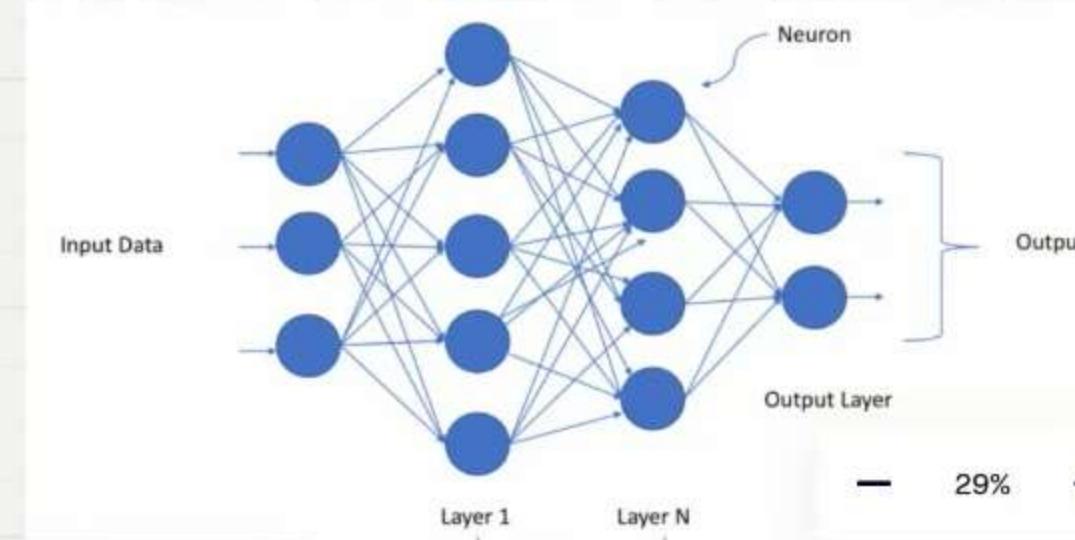


Diagram illustrating a Deep Neural Network structure:

- Input Data**: Represented by arrows pointing into the network.
- Neuron**: Indicated by blue circles representing individual neurons in a layer.
- Output**: Represented by arrows exiting the network.
- Layer 1** and **Layer N**: Labeling the first and last layers of neurons respectively.
- Output Layer**: Specifically labeling the final layer of neurons.

- 29% + ?

Conv x 033b x 1_laD x d415 x ChatC x 1706 x rule-t x Buildi x Heart x CNN x Lab F x vizua x + miro.com/app/board/uXjVKENK1Qc=/ Finish update :

miro | Convolutional Neural Networks | Upgrade | Present | Share

Layer 1 Layer N
Hidden Layer

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

- 37% + ?



③ LLMs vs

can do wide
range of NLP
tasks

Earlier NLP models

designed for specific
tasks like language
translation etc

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

much more!



Conv 033b 1_IaD d415 Chat 1706 rule-t Build Heart CNN Lab P vizua

miro Convolutional Neural Networks Upgrade Present Share

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?

The image shows a Miro board with a light blue background. At the top, there's a toolbar with icons for star, red pen, green square, yellow circle, blue triangle, black square, document, music, download, and a person icon labeled 'R'. To the right of the person icon is a 'Finish update' button. Below the toolbar, the word 'miro' is written in a dark blue font, followed by 'Convolutional Neural Networks' and an 'Upgrade' button. On the far right of the toolbar are 'Present' and 'Share' buttons. The main area of the board has two dashed boxes. The left dashed box contains the handwritten text 'Range of NLP tasks' with a large curly brace underneath. The right dashed box contains the handwritten text 'tasks like language translation etc' with a horizontal underline. Below these boxes, there's a large block of handwritten text in green ink: '"Earlier language models could not write an email from custom instructions, a task that is trivial for modern LLMs"'. At the bottom left, there's a small circular profile picture of a man with glasses and a beard. The bottom right corner features a zoom control with '- 44% +' and a question mark icon. The overall style is a mix of digital interface elements and handwritten notes.

miro | Convolutional Neural Networks | Upgrade | Present | Share

miro.com/app/board/uXjVKENK1Qc=/

“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)

Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward

Forward

Conv x 033b x 1_laD x d415 x Chat x 1706 x rule-b x Build x Heart x CNN x Lab F x vizua x +

miro Convolutional Neural Networks Upgrade Present Share

miro.com/app/board/uXjVKENK1Qc=/

Transformer Architecture (What does it mean)

(This is the only Transformer)

Finish update

44%

Conv x 033b x 1_IaD x d415 x Chat x 1706. x rule-t x Build x Heart x CNN x Lab P x vizua x + miro.com/app/board/uXjVKENK1Qc=/ Finish update :

miro Convolutional Neural Networks Upgrade Present Share

Transformer



It actually something this)

The diagram illustrates the Transformer architecture. It starts with 'Inputs' at the bottom, which are processed by 'Input Embedding'. This is followed by two parallel stacks of layers, each labeled 'Nx'. The first stack consists of 'Multi-Head Attention' (orange box), 'Add & Norm' (green box), and 'Feed Forward' (blue box). The second stack consists of 'Masked Multi-Head Attention' (orange box), 'Add & Norm' (green box), and 'Feed Forward' (blue box). The outputs from both stacks are combined via 'Add & Norm' and then passed through 'Output Embedding' to produce 'Outputs (shifted right)'. Finally, the sequence ends with 'Positional Encoding' and a large 'Add & Norm' block before reaching the 'Softmax' layer, which produces the 'Output Probabilities'.



1



2



3



03762v7 [cs.CL] 2 Aug 2023

Provided proper attribution is provided, Google hereby grants permission to reproduce the tables and figures in this paper solely for use in journalistic or scholarly works.

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

Lukasz 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, based on attention mechanisms, designed explicitly with a global architecture in mind. Our experiments on English-Mandarin translation show that a Transformer trained on back-translation can reach the level of the top-seeded methods on the IWSLT14 English-to-Mandarin task, without any manual tuning of hyperparameters or data processing steps.

Conv x 033b x 1_laD x d415 x Chat x 1706 x rule-k x Build x Heart x CNN x Lab F x vizua x +

miro Convolutional Neural Networks Upgrade Present Share

miro.com/app/board/uXjVKENK1Qc=/

(What does it mean)

Architecture

(This is the only Transformer I know)

It som

The diagram illustrates the architecture of a Transformer, showing its internal components and flow. It starts with 'Inputs' which are processed by 'Input Embedding'. These are then combined with 'Positional Encoding'. The resulting sequence passes through a stack of N_x layers. Each layer contains two parallel sub-layers: 'Multi-Head Attention' (in orange) and 'Feed Forward' (in blue). The outputs from these sub-layers are combined via 'Add & Norm' operations. The final output is processed by 'Output Embedding' and 'Positional Encoding' before being converted into 'Outputs (shifted right)'. A large blue arrow points downwards at the bottom of the board.

Conv x 033b x 1_laD x d415 x Chat x 1706 x rule-b x Build x Heart x CNN x Lab P x vizua x + miro.com/app/board/uXjVKENK1Qc=/

Finish update :

miro | Convolutional Neural Networks | Upgrade | Attention | Attention | Present | Share

(This is the only Transformer I know)

Encoding

Input Embedding

Inputs

Output Embedding

Outputs (shifted right)

A

D

T

C

P

W

A

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

15) 1IM vs Gen AI vs Deep learning vs Machine learning

The Miro board has a title bar with various tabs and a toolbar with icons for selection, zoom, and sharing. The main area contains handwritten text in pink and blue ink. A large pink arrow points down from the top text to a dashed-line box containing blue text. A red box highlights the word 'Transformers' in the blue text. To the right of the text is a diagram of a Transformer's input and output paths. The input path shows 'Inputs' leading to 'Input Embedding' and then to 'Encoding'. The output path shows 'Outputs (shifted right)' leading to 'Output Embedding' and then to 'Encoding'. A red arrow points from the handwritten text 'I know)' towards the output embedding stage.

Conv x 033b x 1_laD x d415 x Chat x 1706 x rule-t x Build x Heart x CNN x Lab F x vizua x + miro.com/app/board/uXjVKENK1Qc=/ Finish update :

miro | Convolutional Neural Networks | Upgrade | Present | Share

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

⑤ LLM vs Gen AI vs Deep learning vs Machine learning

Large language Models (LLM)

Deep learning (DL)

Machine Learning (ML)

(AI)

The image shows a Miro board with a grid background. At the top, there's a toolbar with various icons for editing and sharing. On the left, a vertical sidebar contains icons for text, shapes, and other tools. The main area has handwritten text and diagrams.

- Text:** "Don't worry. We will learn all about this secret sauce: **Transformers**". A large blue arrow points down to the word "Transformers".
- Text:** "⑤ LLM vs Gen AI vs Deep learning vs Machine learning". Red arrows point from the numbers ⑤, 1, 2, and 3 to the words LLM, Gen AI, Deep learning, and Machine learning respectively.
- Diagram:** A large dashed circle encloses the four learning paradigms. Inside this circle, another dashed circle encloses the first two: "Large language Models (LLM)" and "Deep learning (DL)". Below this, another dashed circle encloses "Machine Learning (ML)".
- Text:** "(AI)" is written at the bottom right of the main text area.

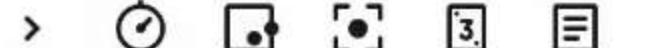
Bottom right corner: - 24% + ?

miro

Convolutional Neural Networks



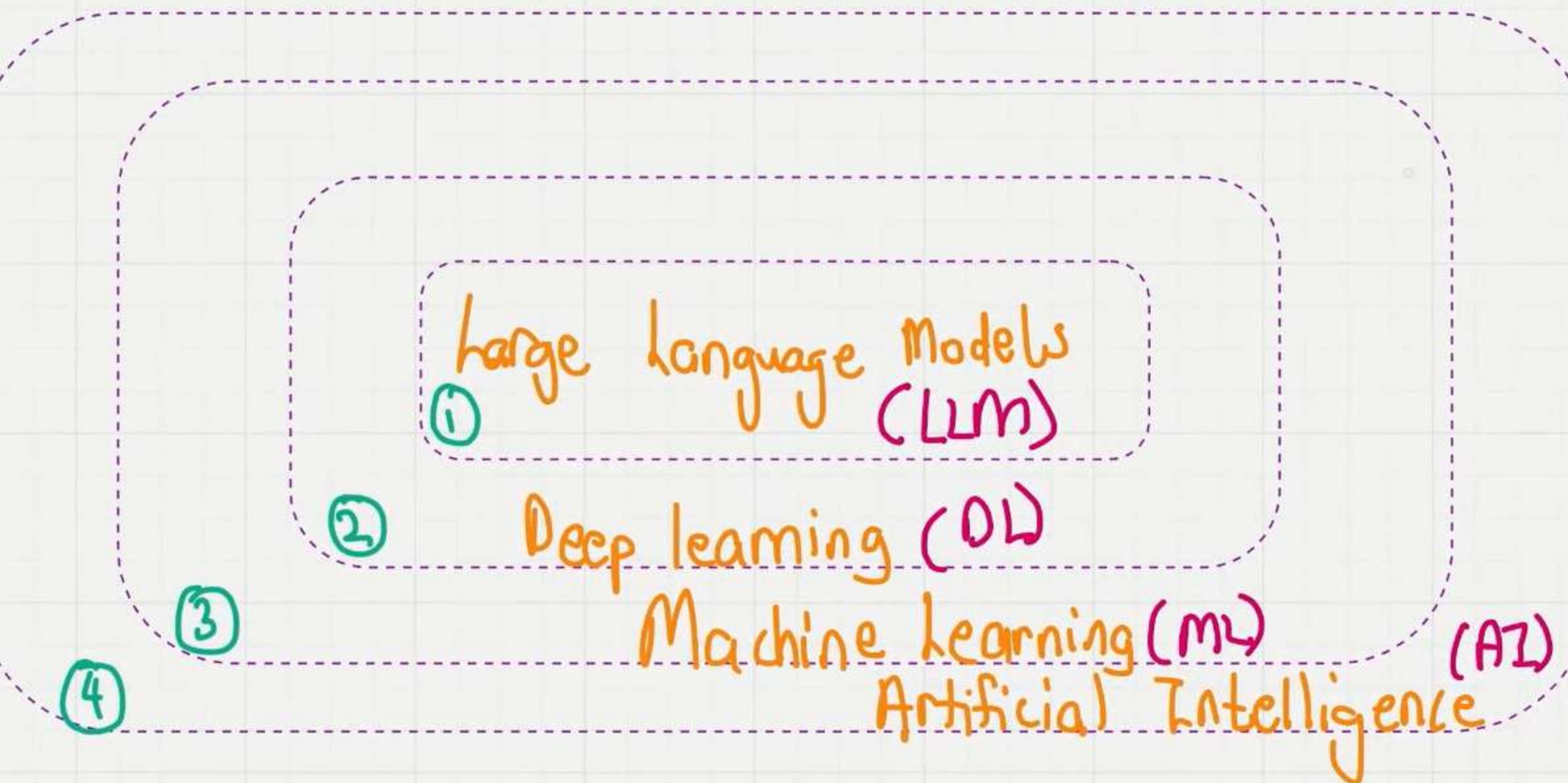
Upgrade

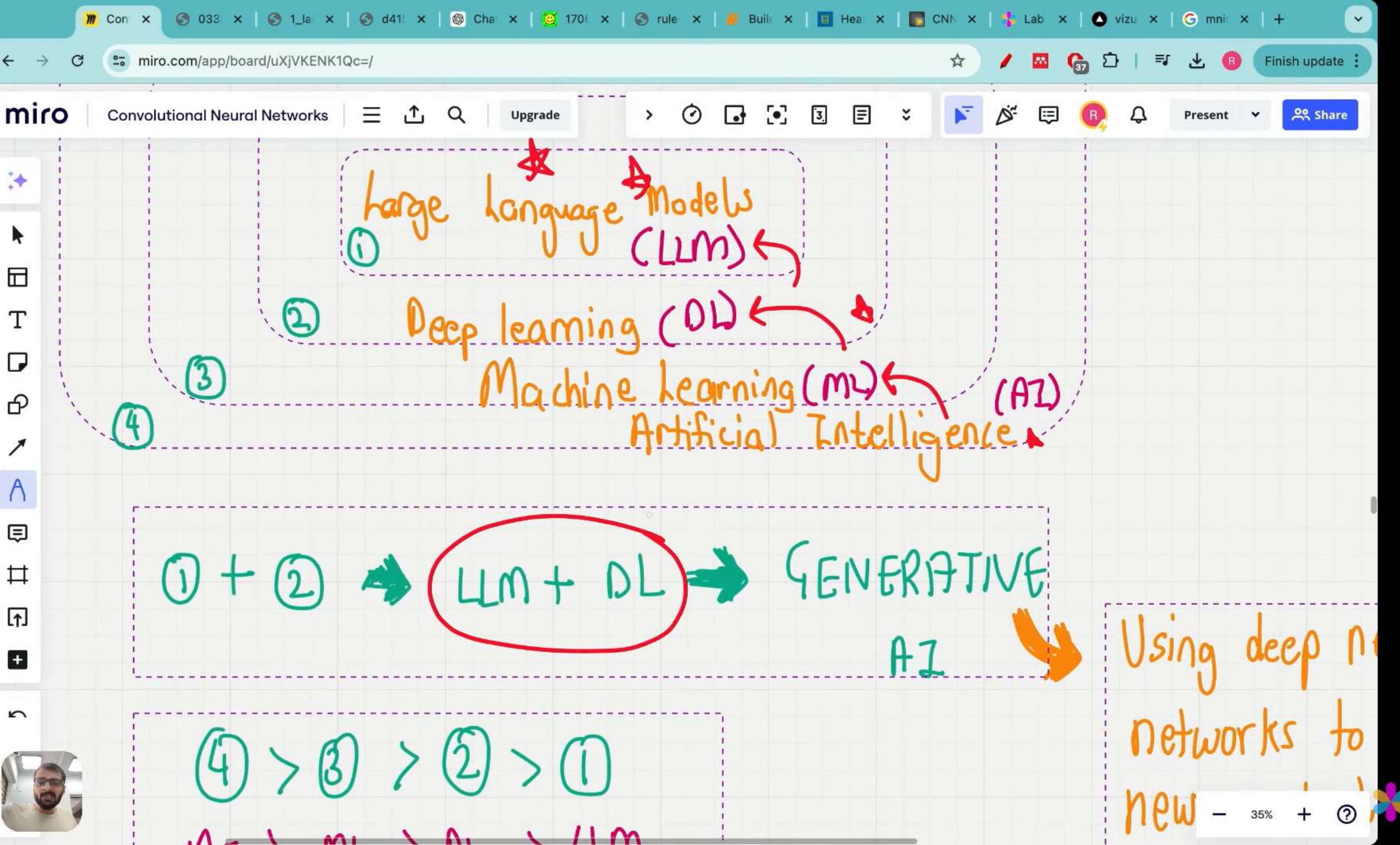


Present

Share

LLM vs Gen AI vs Deep learning vs Machine learning





miro Convolutional Neural Networks Upgrade Present Share

Finish update

Machine Learning (ML) + Deep Learning (DL) = Artificial Intelligence

LLM + DL → GENERATIVE AI

② > ①
DL > LLM

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

get a specific application task done

Con x 033 x 1_la x d41 x Cha x 170 x rule x Built x Hea x CNN x Lab x vizu x G mni x +

miro Convolutional Neural Networks Upgrade Present Share

large language Models (LLM) ←
① Deep learning (DL) ←
② Machine learning (ML) ←
③ Artificial Intelligence (AI)
④

① + ② → LLM + DL → GENERATIVE AI

④ > ③ > ② > ①
AI > ML > DL > LLM

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

LLMs represent a specific application of deep learning techniques leveraging

- 23% + ?

Con x 033 x 1_la x d41 x Cha x 170 x rule x Built x Hea x CNN x Lab x vizu x G mni x +

miro | Convolutional Neural Networks | Upgrade Present Share

Finish update : 37

④ > ③ > ② > ①

AI > ML > DL > LLM

networks to create new content such as images, various

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

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

Youtube Generator
Generate guiding questions aligned to a YouTube video.

Text Question
Generate text-dependent questions for students based on any text that you input.

Worksheet Generator
Generate a worksheet based on any topic or text.

MCQ Generator
Create a multiple choice assessment based on any topic, standard(s), or criterial.

Text summarizer
Take any text and summarize it in whatever length you choose.

Text Rewriter
Take any text and rewrite it with custom criteria however you'd like!

Proof Read
Take any text and have it proofread, correcting grammar, spelling, punctuation and adding clarity.

Lesson Plan
Generate a lesson plan for a topic or objective you're teaching.

Report Card
Generate report card comments with a student's strengths and areas for growth.

- 28% + ?

Con x 033 x 1_la x d41 x Cha x 170 x rule x Built x Hea x CNN x Lab x vizu x G mni x +

miro Convolutional Neural Networks Upgrade

human like text J

Applications of LLMs

① Content creation

② Chatbots/
Virtual
Assistants

③ Machine
translation

④ novel text
generation

⑤ Sentiment
analysis

⑥

Chatbots/
Virtual
Assistants

Machine
translation

novel text
generation

Sentiment
analysis

Content creation

Youtube Generator

Text Question

Worksheet Generator

MCQ Generator

Text summarizer

Text Rewriter

Proof Read

Lesson Plan

Report Card

Essay Grader

PPT Generator

Grade Essay

Generate PPT

26% - + ?

Finish update :

Con x 033 x 1_la x d41 x Cha x 170 x rule x Buil x Hea x CNN x Lab x vizu x G mni x +

miro.com/app/board/uXjVKENK1Qc=/

miro Convolutional Neural Networks Upgrade Present Share

Assistants

VISUALIZATION GENERATION MAPPING

Youtube Generator Text Question Worksheet Generator

MCQ Generator Text summarizer Text Rewriter

Proof Read Lesson Plan Report Card

Essay Grader PPT Generator

“The sky is the limit when it comes to LLM applications”

Finish update

26%

A

Lecture 3 : Stages of building LLMs

Creating an LLM = Pretraining + Finetuning

Training on a large, diverse dataset

Refinement

training on domain-specific data

Lecture 3: Stages of building LLMs

Creating an LLM = Pretraining + Finetuning

Training on a large, diverse dataset

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

Pretraining + finetuning schematic:

Pretrained L

Proportion in training data

A %

A %

A %

A %

A %

A %

A %

A %

Con x 200 x Imp x vizu x Cha x Cha x Intro x Gen x BS JPM x Intro x lang x com x Ope x +

miro.com/app/board/uXjVKENK1Qc=/

Convolutional Neural Networks | Upgrade Present Share

Dataset name Dataset description Number of tokens Proportion in training data

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 +

- 27% + ?

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

Pre-training cost for GPT-3

27% + ?



Research

Products

Safety

Company

pervised learning?

ks

: Dataset examples

● Sentiment Analysis— Pre-trained● Winograd Schema Resolution----- Random Init● Linguistic Acceptability● Question Answering

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



miro

Convolutional Neural Networks



Upgrade



Present

Share

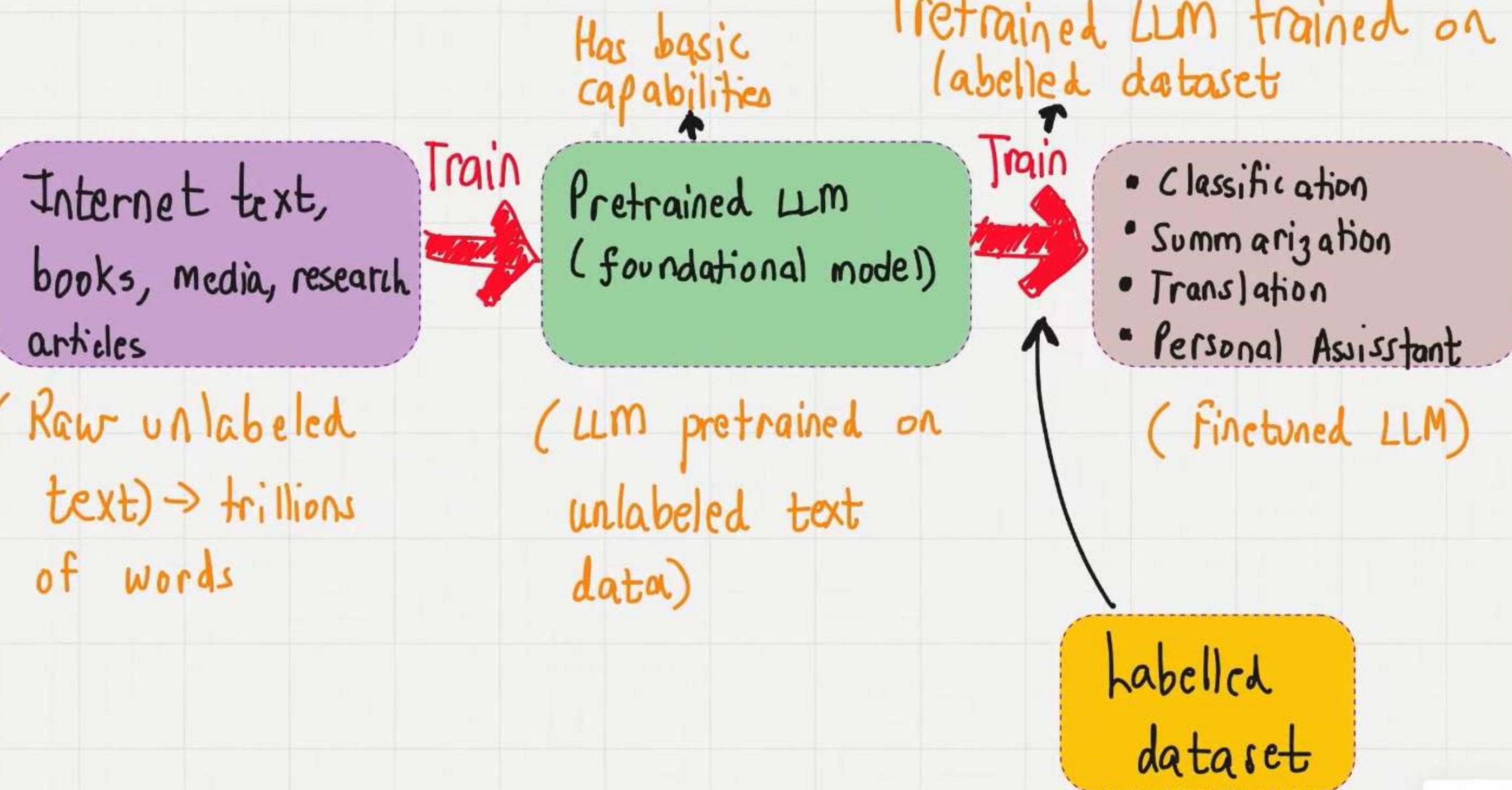
= Pretraining + Finetuning

ining on a
ge, diverse dataset

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



Pretaining + finetuning schematic:



m 200 x | vizi x | Cha x | Cha x | Intro x | H Gen x | BS JPM x | Intro x | lang x | G com x | Ope x | +

miro Convolutional Neural Networks Upgrade Present Share

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

•

↶

☒

T

☐

🔗

A

💬

#

↑

+

↓



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.

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.

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



miro

Convolutional Neural Networks



Upgrade



Present



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
 - finetuning for classification tasks

labeled dataset consists of



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

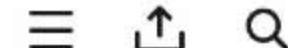
finetuning for
classification tasks

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



miro

Convolutional Neural Networks



Upgrade

category or ...

instruction finetuning

finetuning for
classification tasks

A labeled dataset consists of

instruction-answer pairs.

text translation,

airline customer support

labeled dataset

consists of text &

associated labels.

eg: emails → spam vs no-spam.

