

< DeepLearning, AI & UPSTAGE AI with Andrew NG >

* Multi-stage Continued Training

Pre-trained model → Fine-Tuned Model - Aligned Model

In usual case: Pre-trained → Final Tuned alt

↳ use in specific domain
or

use their own data in

Fine-Tuned Model & Aligned Model

exceptional: Fine-Tuned & Aligned Model changing
(but quite often)

⇓

Not enough.

should change or give some dataset in
pretrained model

↳

usually occur in new domain.

(not just changing topic, contents)

↙

but: Same weights (parameters): Cost & Time & Issues: Pre-trained model
was most expensive.

< Dataset for Pre-Training >

Kind of dataset: Unstructured text → ex) series of books, articles, whatever

↳ keep predicting next word

↳ while this phase: weights are updated

< Dataset for fine-tuning >

Kind of dataset: highly structured text

ex) question - answer pairs, instruction - response pairs etc.

↳ Pre-Training: Reading books)

~~~~~  
Fine-Tuning: Testing, exams

↳ data should be very specific and precise.

Traditionally, performed by human. but right now, trying LLM to generate dataset.

xxxxxx Quality of data.

↳ concludes the performance of LLM.



Data cleaning → Required. → Method:



From UPSCALE AI - Data verse - open source AI

↙  
Data-cleaning - SW.

1. Deduplication.

duplicate data could bias the model  
in specific pattern.

2. Quality Filters

ex) Focusing on specific language.

3. Contents Filters

→ No more bad words (don't let bias to toxic contents)

4. Privacy Reduction & Rule-based cleaning

↳ Prevent data leakage

↓  
revise poorly formatted  
text

## [ Data - cleaning ]

Form of

1. web crawling → Pre Training dataset.

2. Structured data → Unstructured data ( ∵ Pre - Training Model )

3. Cut out rows which is too short ( Under 3 words )

4. Paragraphs - duplicates /  $\text{len}(\text{paragraph}) > 0.3$  ∴ Remove

↳ deduplication.

cf) Using HuggingFace : We can use quite fine datasets.

↳ have done Preprocessing a lot.

5. language filter → specify to specific language.

cf) Parquet : column storage file format

used in big data , big data analysis

( kind of CSV )

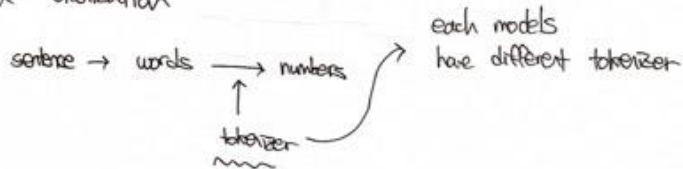
xxxxx

[Data Packaging : Tokenizing + Packing]

Tokenizing : breaking each <sup>text</sup> into smaller meaningful units, which are called "tokens"

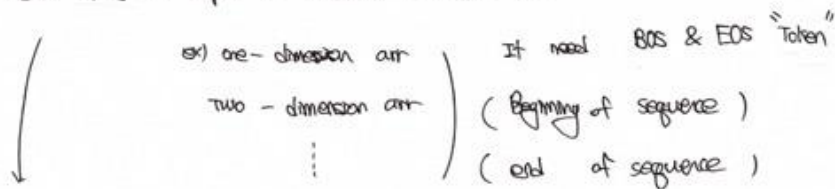
Packing : "Packing" tokens into the maximum sequence length to improve training efficiency

xxx Tokenization



xxx Packing

each blocks : shaped in different kind of AS



Training LLM : same length of data is much efficient.

~~xxxx~~ Repurpose model for training

This section was about the way to customize, generate the one LLM model.

cf) In this course, only use (focus) on "decoder-only - Transformer"

using this model, we'll predict the text and let the model operate autoregressive model properly.

autoregressive model: making output depends on its own previous value (dataset)

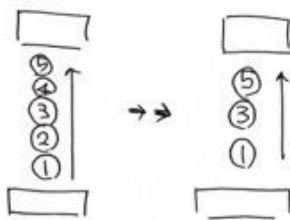
xx decoder only Transformer model

: text  $\rightarrow$  vector  $\rightarrow$  weights layer

making & Training Model: with random weights: too much time & cost

$\rightarrow$  let's just use existing open-source sw's weights.

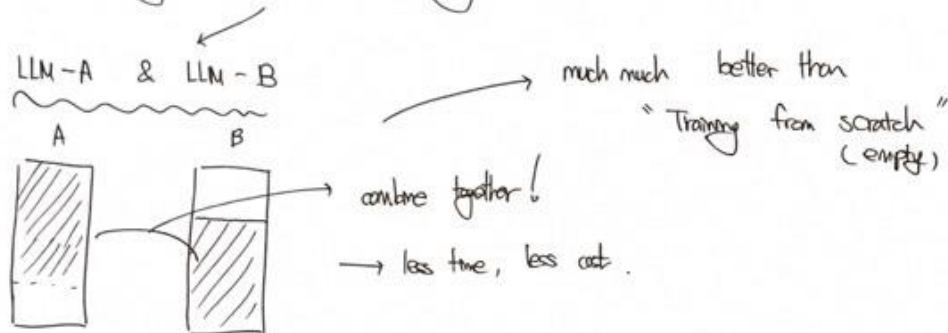
~~xxxx~~ customizing "Downscaling"  
LLMs with



: removing some layers from

Big LLM. But in small models, it doesn't work well.

## Customizing LLMs with Upscaling



## c) About "Model Size"

1. Number of Parameters = weight + bias ...  
 ↳ trained Parameters.
2. Architecture = number of layers, number of nodes, and the way they connected.  
 ↳ deep neural network > shallow neural network.
3. Precision  $\Rightarrow$  model use.  
 ↳ the number of bits floating point.  
 ex) 32-bit floating point.  
 in "model size" (usually)
4. Compression: retain the model's performance  
 Techniques. but minimize the model size  
 ex) pruning, quantization, knowledge distillation.

mobile phone, embedded system: need smaller model.

↳ ∴ model size is important component.

## < Training Cycle >

\* 1. Data Prep

2. Hyper parameter configuration

3. Training

4. Monitoring

a lot of time  
would be taken.

cf) you can use  
Training Cluster to check  
how it would be taken.

Common Benchmark Dataset (Evaluating ML/DL model)

ARC, MMLU, HellaSwag, TruthfulQA

WinoGrande, GSM8K, MT Bench, ELI Bench

IFEval

They evaluate specific part of model  
all

ex) Language understanding, True or false, Mathematical Reasoning ...