(5) Transformers vs LLMs

- Not all transformers are LLMs
- Transformers can also be used for computer vision

vision transformer → can be used for image classification, like CNN

- Not all LLMs are transformers
- can be ~~LSST~~ LSTM or recurrent NN

Closer look at GPT

(1) zero-shot vs few shot learning

→ zero-shot = Generalize to completely unseen tasks w/ out prior specific examples

- few-shot → Learning from a minimum number of examples which user provides as input

- Generative pre-training (GPT)
  - text used is not labelled
  - don't need to provide labels
  - Generating next word
  - Transformers + unsupervised pretraining

(OpenAI)

- GPT 3 1.75B parameters
-

Zero-shot: model predicts answer given only a natural language description of task. No gradient updates are performed

Translate english to french (desc) GPT-4 can do
cheese -> (prompt) zero-shot ted

one-shot: in addition to task description, model sees a single example of the task. No gradient updates are performed

Translate English -> French: (task desc)
Sea otter => Loutre de mer (example)
cheese -> (prompt)

Few-shot: sees a few examples GPT-4 is better w/ few shot

seaotter -> loutre de mer
peppermint -> menthe poivree
plush giraffe -> girafe peluche
} examples

cheese -> GPT-3 is few shot learner
give few examples for translation

Token -> unit of text which model reads for many, can think of token = 1 word

Total pre-training cost for GPT-3 is $4,6 million

* pretrained models are base/foundational models which can be used for further finetuning

* Many pretrained LLM's are available as open-source models -> can be used as general purpose tools to write, extract and edit text which was not part of training data

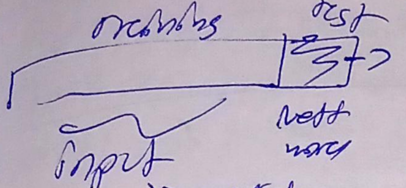
③ GPT Architecture (only decoder block, no encoder)

- simply trained on 'next-word' prediction tasks
  - ↳ the lion roams in the jungle
    next word

* W/ this training, they can do a wide range of other tasks like translation, spelling correction etc. (output)(labels)

ex: second law of robotics: a
second law of robotics: a robot
second law of robotics: a robot must

* Next word prediction: Self-supervised learning
↳ self-labelling



↳ predicted word

Minimize diff between

* we don't collect labels for training ~~data~~, but use structure of data itself
↳ Next word in sentence is used as label } Auto regressive model; use previous outputs as inputs for future predictions

① Pretraining
↓
unsupervised structure of data itself

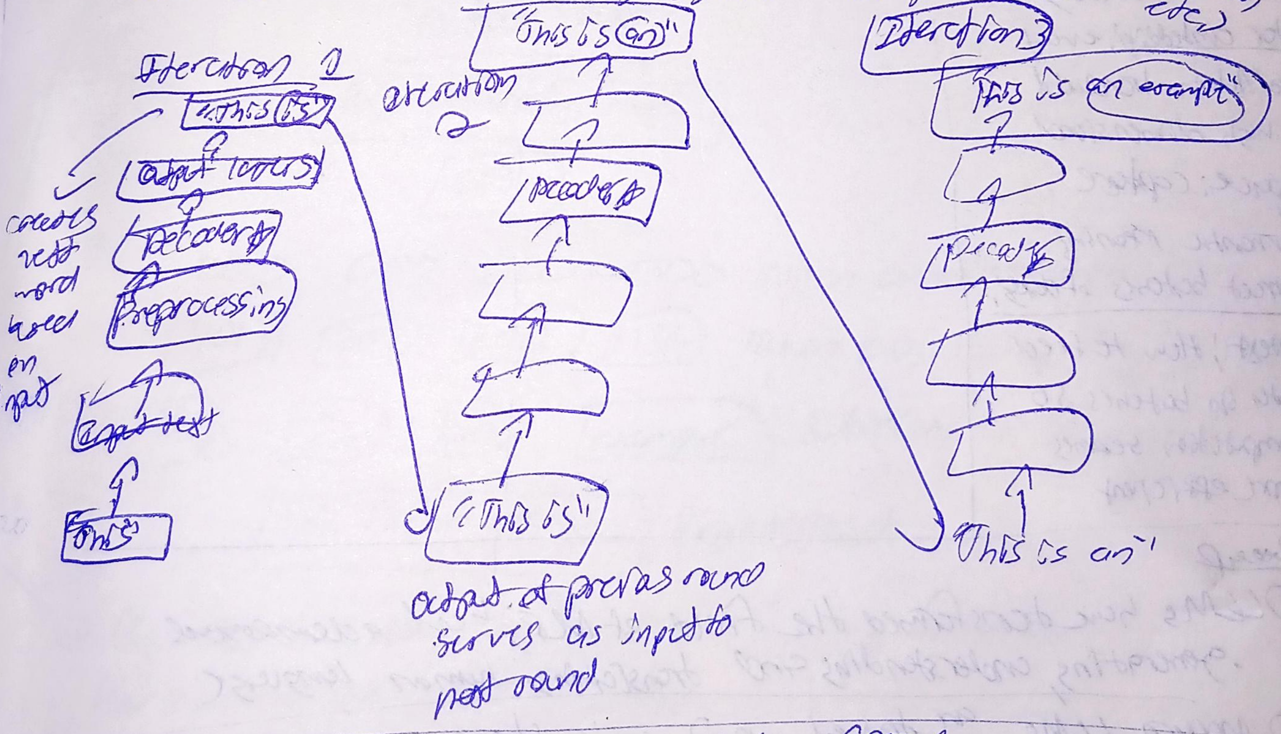
Auto Regressive models

* compared to original transformer architecture, GPT arch is simpler

* There is no encoder, we just have the decoder

Original transformer: 6 encoder-decoder blocks 96 transformer layers

GPT-3: 96 transformation layers, 175 billion parameters (decoder, etc)



output of previous round serves as input to next round

Although trained only for next word prediction, GPT model can perform other tasks like ~~[illegible]~~ translation

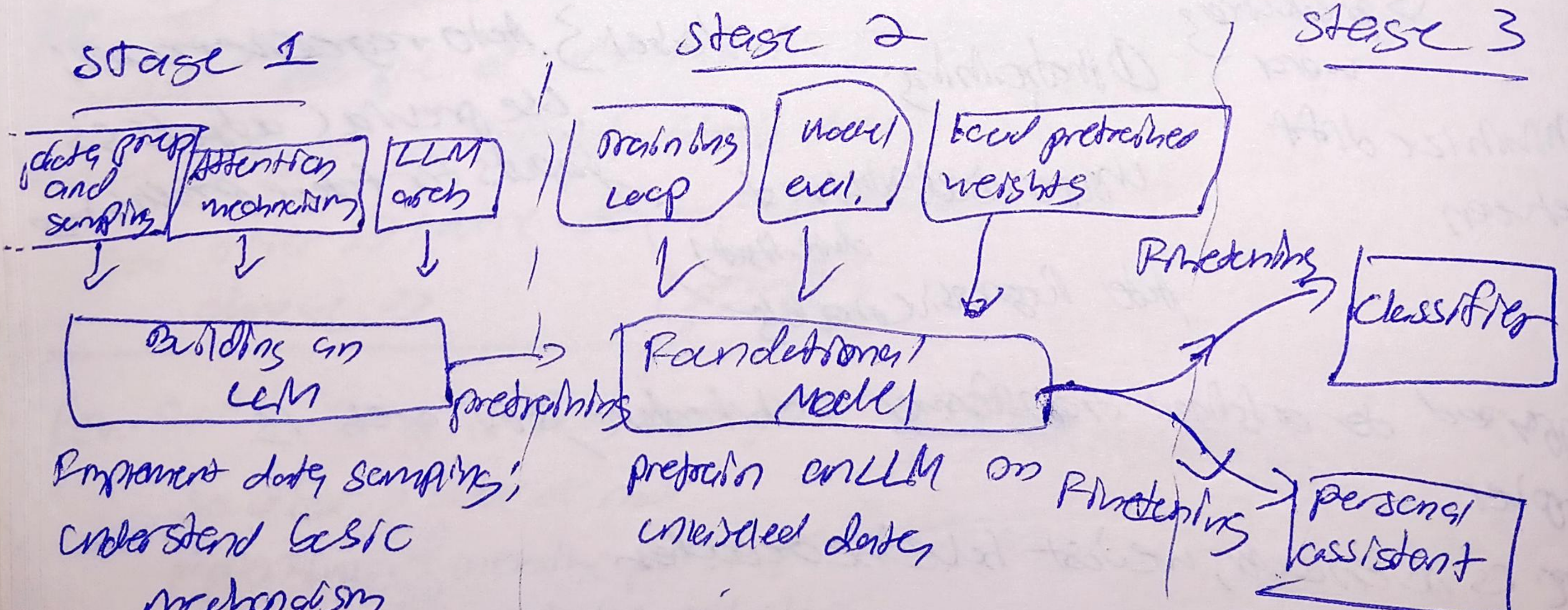
↳ Emergent Behavior

"Emergent Behavior" -> ability of a model to perform tasks that the model wasn't explicitly trained to perform

How to explain? Still isn't known

o Good research topic

Stages of Building LLM



Implement data sampling; understand basic mechanism

pretrain an LLM on unlabeled data

Finetune pretrained LLM

tokenization; break text into tokens

vector embedding event