Lesson 3: Pretraining and finetuning for building LLM

1 token ≈ 1 word lets say. <u>Pretraining (First Stage)</u>

Trained on 410 billion tokens from CommonCrawl

webtext2 -> 19 billion

Books -> 67 billion

wiki -> 3 billion

Different weights.

LLM perform well b/c trained on Huge amounts of data

↳ overcompletion

The lion is in the <u>forest</u>; trained on this type of task (translating)

↳ can do a wide range of tasks as well.

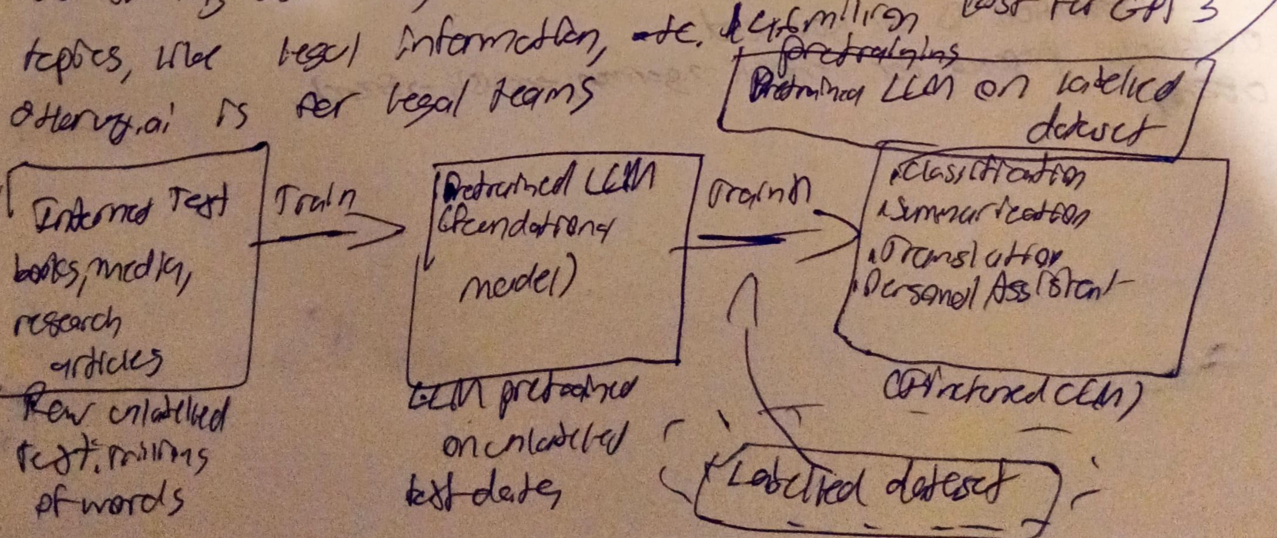
Can begin to perform tasks w/ out ever training on them. (MC, summarizing, etc)

Pretraining -> Train LLM on huge data so can do wide range of tasks

<u>Finetuning</u> (Second Stage)

- CBO airline; you want chatbot so users can ask questions maybe?
  - ↳ you want specific responses, not generic.
  - ~~want~~ be specific to company or application
- High-quality is why ppl finetune; refinement on smaller dataset; Train model on own data.
- OG training data may lack extensive knowledge of certain topics, like legal information, etc. 4.6 million cost for GPT 3
- Harvey.ai is for legal teams

fraud detection


~~pretraining~~
Pretrained LLM on labelled dataset
Internet Text
books, media, research articles
Raw unlabelled text: millions of words
Train
Pretrained LLM (foundational model)
LLM pretrained on unlabelled text data
Train
.Classification
.Summarization
.Translation
.Personal Assistant
(Finetuned LLM)
Labelled dataset

Steps for building LLM

① Train on large corpus of text data (Raw text)

↳ Regular text w/ out any labelling info

② First stage of LLM (pretraining)

◦ GPT-3 model is pretrained model which is capable of text completion

③ After obtaining pretrained LLM, we can further train LLM on labelled data - finetuning

④ Two popular finetuning categories

↙ Instruction fine tuning

↙ Labelled dataset consists of instruction-answer pairs

- text translation, airline customer support

↘ Classification tasks fine tuning

↓ Labelled dataset consists of text and associated labels,

◦ emails → spam vs not spam

---

## Basic intro to transformers

◦ Deep neural network arch. most modern LLMs use

↳ 2017 paper introduced

↳ Attention is all you need

◦ originally proposed for english → german, english → french.

◦ ~~Better~~