Build LLM from scratch (by Sebastian Raschka)

1. Intro to LLM → Neural network designed to understand, generate & respond to human like text

Deep Neural Network

Trained on Massive Amount of data

→ Large Language Models

Large → Models have billions of parameters

Language Models → These models do a wide range of NLP tasks: Question-Answering
↳ translation,
↳ sentiment Analysis &
↳ much more

→ LLM vs NLP model (earlier)

LLM → Can do wide range of NLP tasks

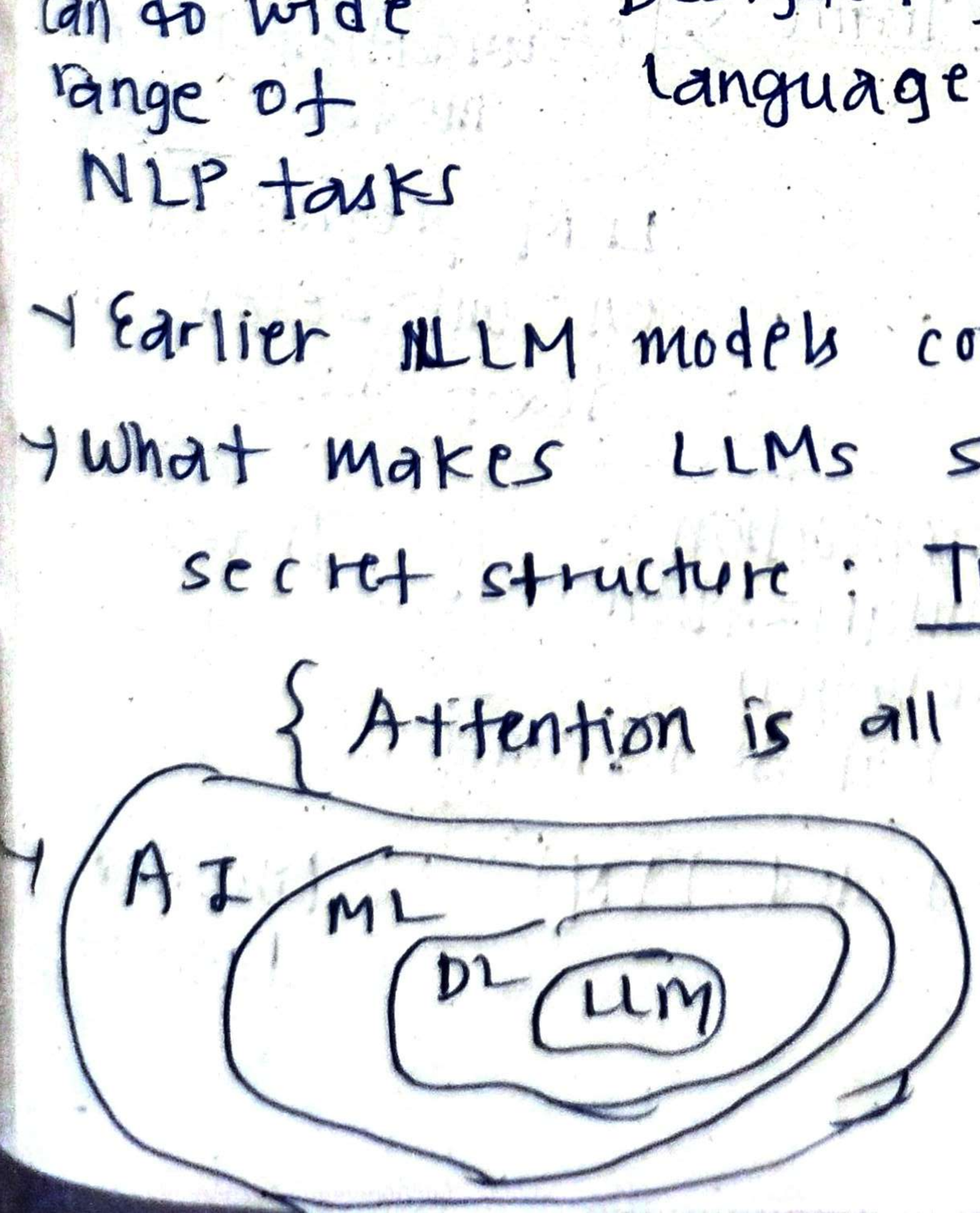
NLP model (earlier) → Designed for specific tasks like language translation etc.

↳ Earlier LLM models couldn't write email.

↳ What makes LLMs so good?

secret structure: Transformer Architecture.

{ Attention is all you need }



AI → Rule Based
ML → learning Based
DL → Neural Netwm

LLM + DL → GenAI.

Applications of LLM :

- content creation
- Chatbots / Virtual Assisstants
- machine Translation
- Novel text generation
- sentiment Analysis

The sky is the limit for LLM Application

→ stages of Building LLMs

Creating an LLM = Pretraining + Finetuning

Train on a large, diverse dataset

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

Pretraining + Finetuning schematic

Data
Internet texts, books, media research, articles

Raw unlabled text
+Trillions of words

Train →

Has basic Cababilities

pretrained LLM (foundational model)

LLM pretrained on unlabeled text data

Train ←

When pretrained LLM trained on labelled data

- Classification
- summarization
- Translation
- Personal Assisstant

(Finetuned LLM)

labelled dataset

*steps for building a LLM:

① Train on a large corpus of text data (raw text)
↳ Regular text without any labeling information.

② First training stage of LLM is also called pretraining.
↳ creating an intial pretained LLM (base / foundational model)

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

← 2 types

Instruction Fine tuning

Labelled dataset ← Consists of instruction-answer pair.

e.g. text translation, airline customer support

Finetuning for classification tasks

Labeled dataset Consists of texts & associated labels.

e.g, email
↳ spam
↳ Nospam.