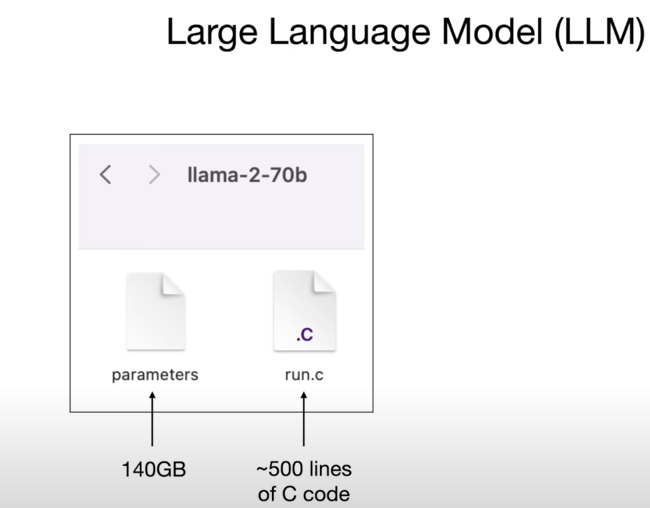
LLM - basically 2 files in a directory - parameters , run.c



Llama 270b model – LLM released by meta api

Llama-2-70b: Llama series of language model , 2nd iteration , 70b parameters

There are other llama-2 series model with 7b, 13b, 34b and 70b

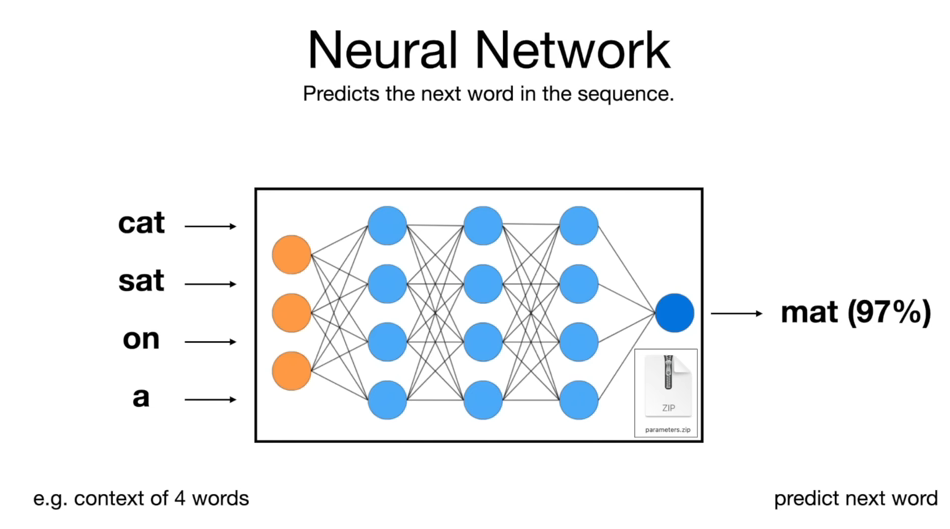
Every parameter is stored as 2 bytes(float 16), so 70billion\*2bytes = 140GB (70\*10^9\*2\*8 bit in gb = 70\*2 GB = 140GB )

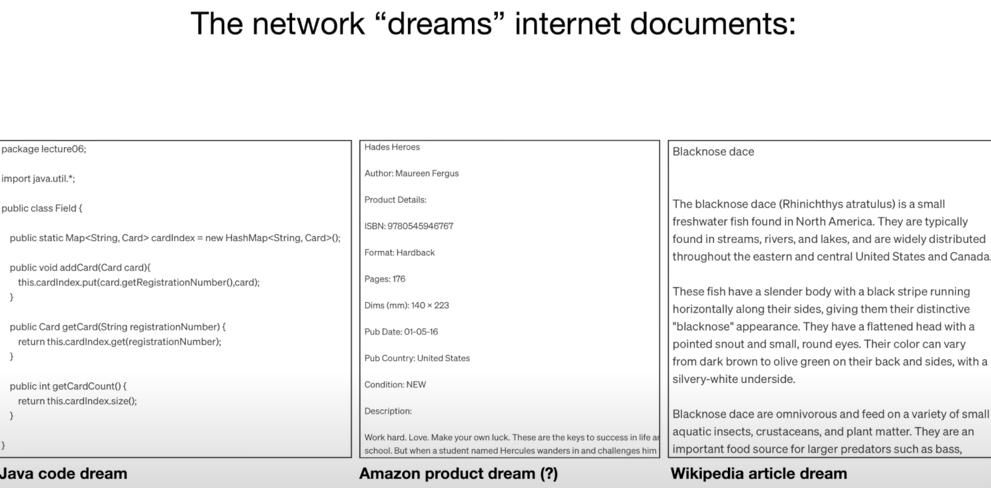
**Run.c** – program that run the neural network. In C it requires only 500 lines of code. When compiled, generates a binary that points to the parameters file. The code basically has well established algorithms and runs a neural network.

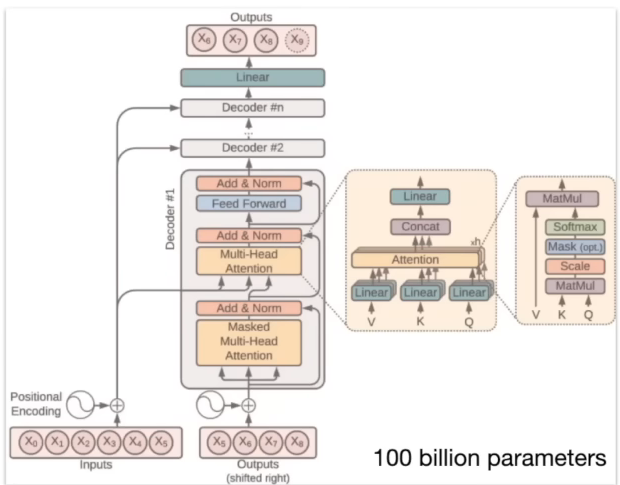
**Parameter**- AKA weights of neural network.

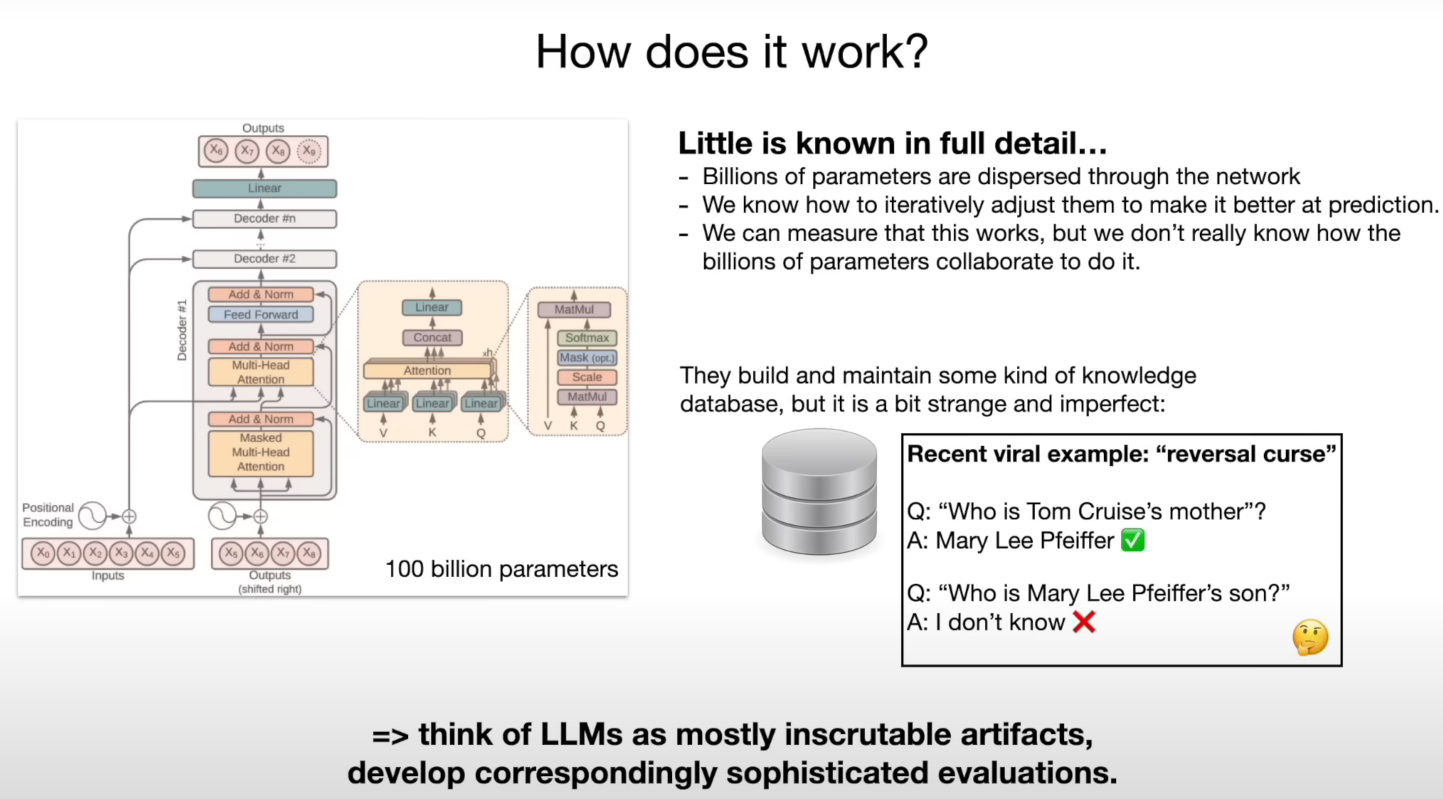
Taking these two files, we can run them on a pc/Mac book. (This is called inference)

Parameter are obtained by the process of ‘training the model’. It’s much more involved process than that of inference. Think of it as a compression of a good chunk of internet. ~ 10TB of text. It is obtained by crawl of the internet. Takes 600 GPUs and costs $2milliion. Compression is lossy compression unlike zip compression which is reversable.



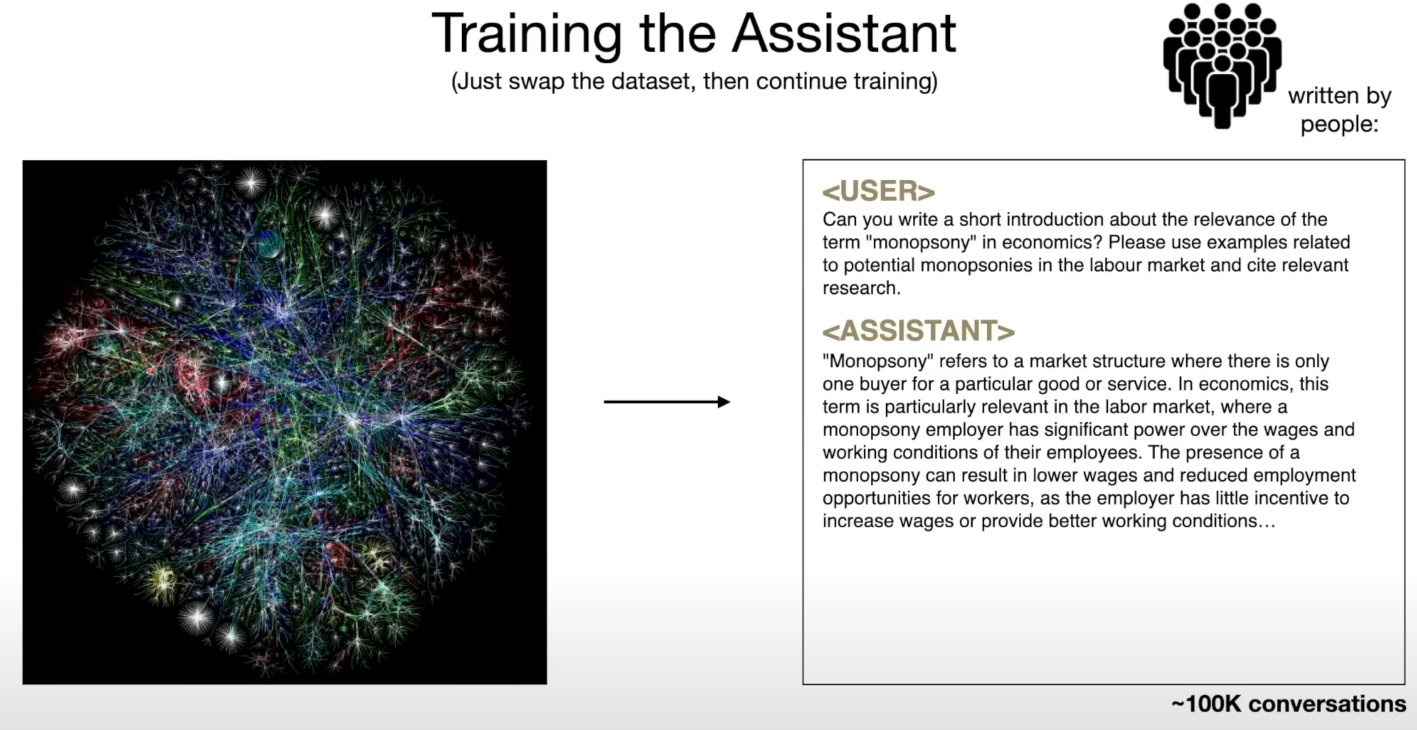




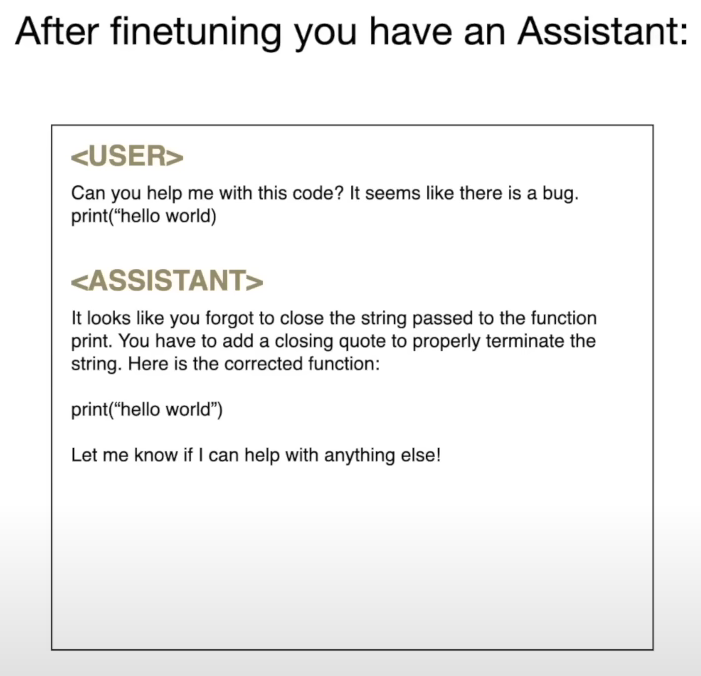


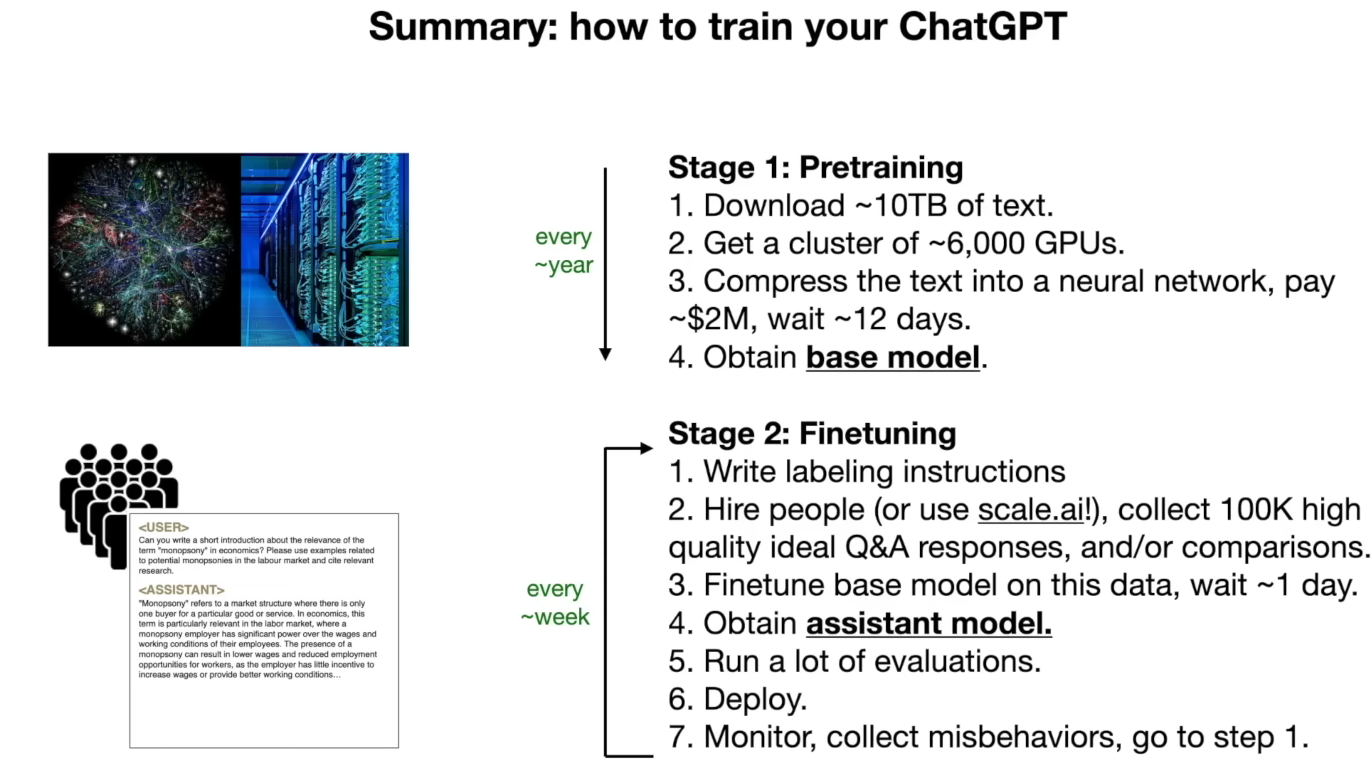
This is first stage of training called pre-training (internet document generation/next word prediction)

Second stage of training is fine tuning. In this stage we obtain “assistant model”.



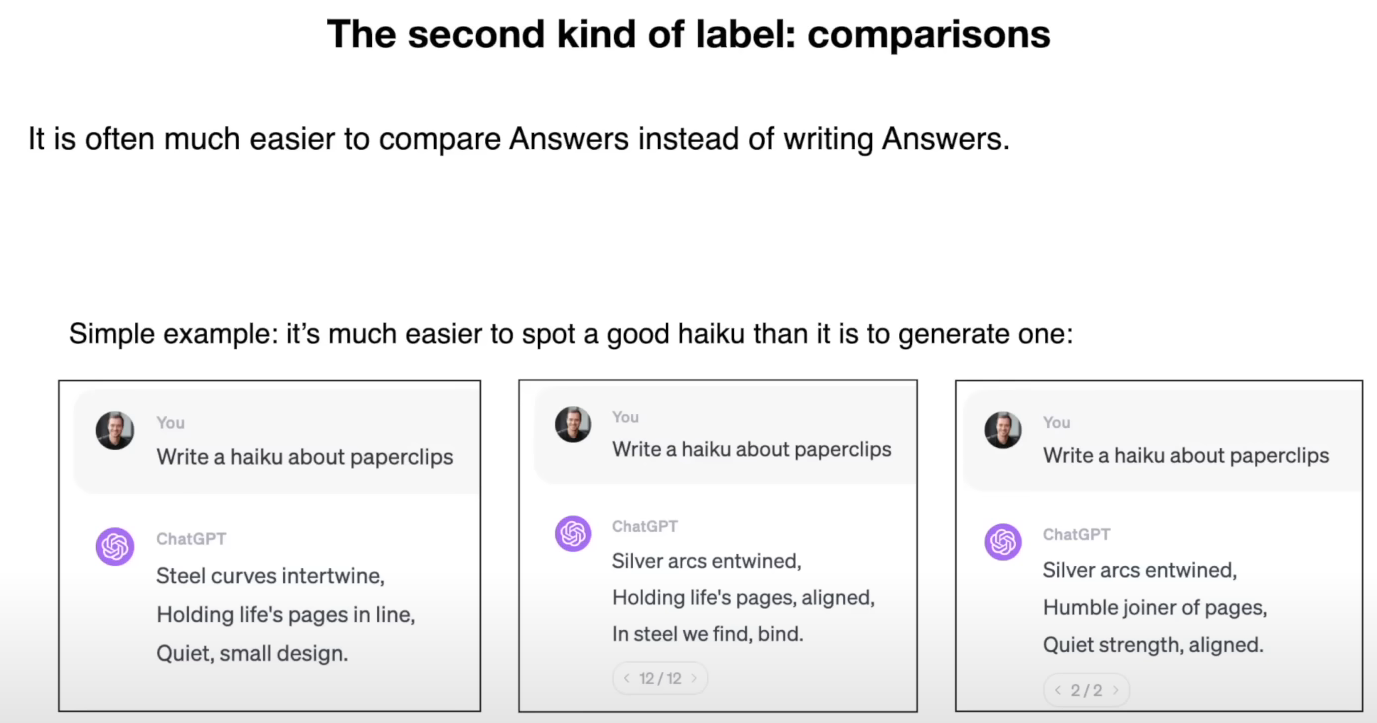
In this stage instead of internet sourced text, documents are used which are manually written documents using labelling instruction with questions and expected ideal response. So basically dataset is swapped with these high quality Q&A documents. This stage is about changing the formatting from internet documents to q&a documents. After this stage we get an Assistant.





Llama-2 series comes with both base model as well as assistant model.

Optional stage 3- comparison based fine tuning



RLHF- reinforced learning from human f…

Example of labelling instructions-