LLM BASICS & FLOW

So far:

1) Intro to AI / ML - NNs
2) ML model foundations
3) LLM foundations

Large Language Model

- Large: big scale { 100B - 5T params } Pre-trained on internet-scale data
- Language: trained on text for most languages in the world
- Model: ML model, specifically, NN

GPTs

- Generative: generates text, image, video, audio
- Pre-trained: already trained on large datasets of text, audio, video, images
- Transformers: a type of NN architecture pioneered first by Google in 2017 ("Attention is all you need")

OPEN-SOURCE	CLOSED (COMMERCIAL)
Meta – Llama 2/3 (3B / 70B / 7B)	openAI – GPT3.5, GPT4
BERT	Azme – GPT3.5, GPT4
Code Llama	Anthropic – Claude 3?
Stability – STABLE LM, STABLE Image	hoogle – Gemini Pro
MISTRAL 8x7B (45B)	
HuggingFace → marketplace of sorts for open-source AI LLMs in particular	

default cloud provider
AWS, Azure,
1)
2) deep Infra
3) BedRock

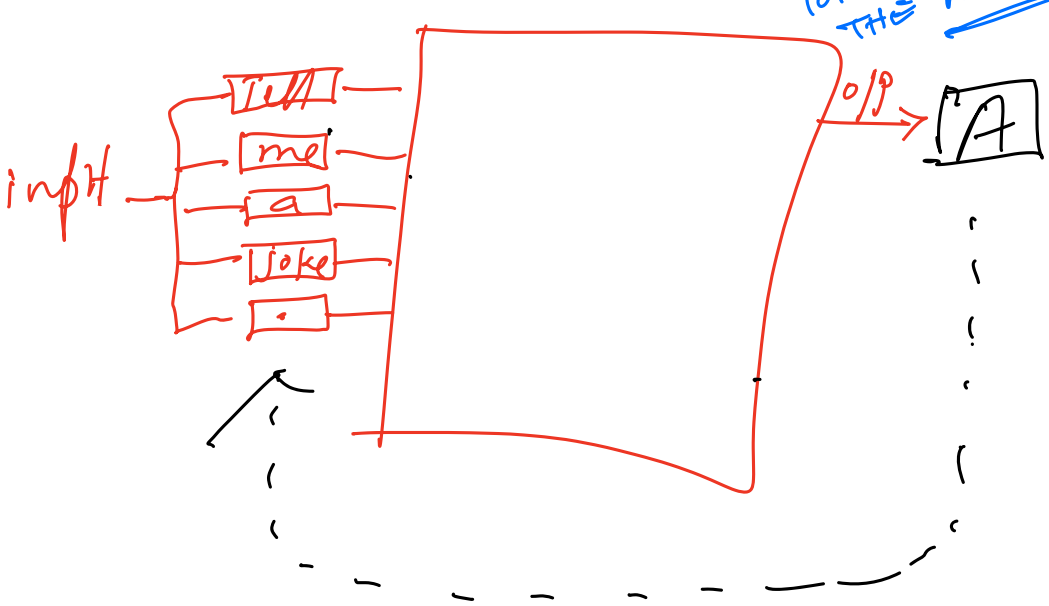
How do LLMs go through generating text:

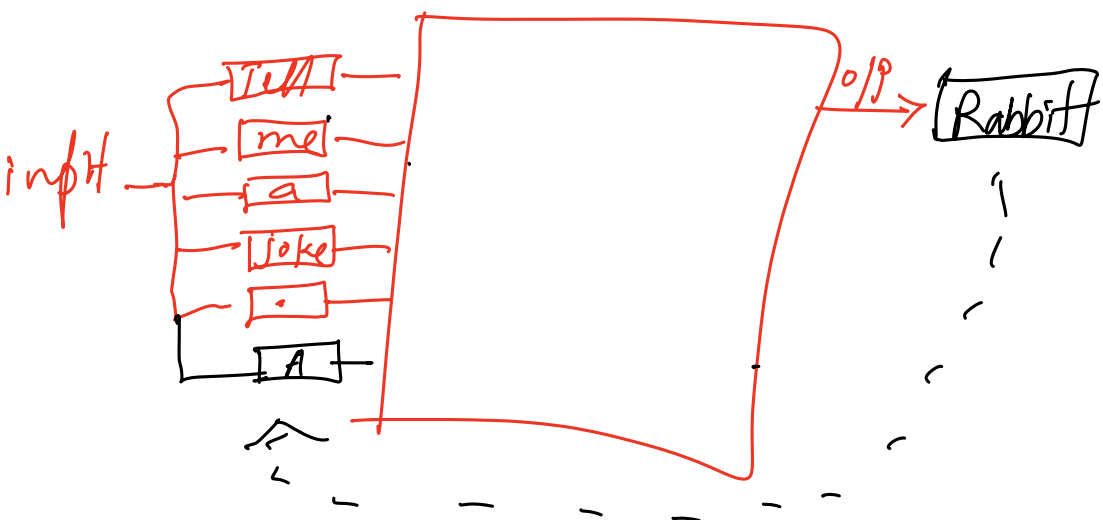
Prompt: Tell me a joke.

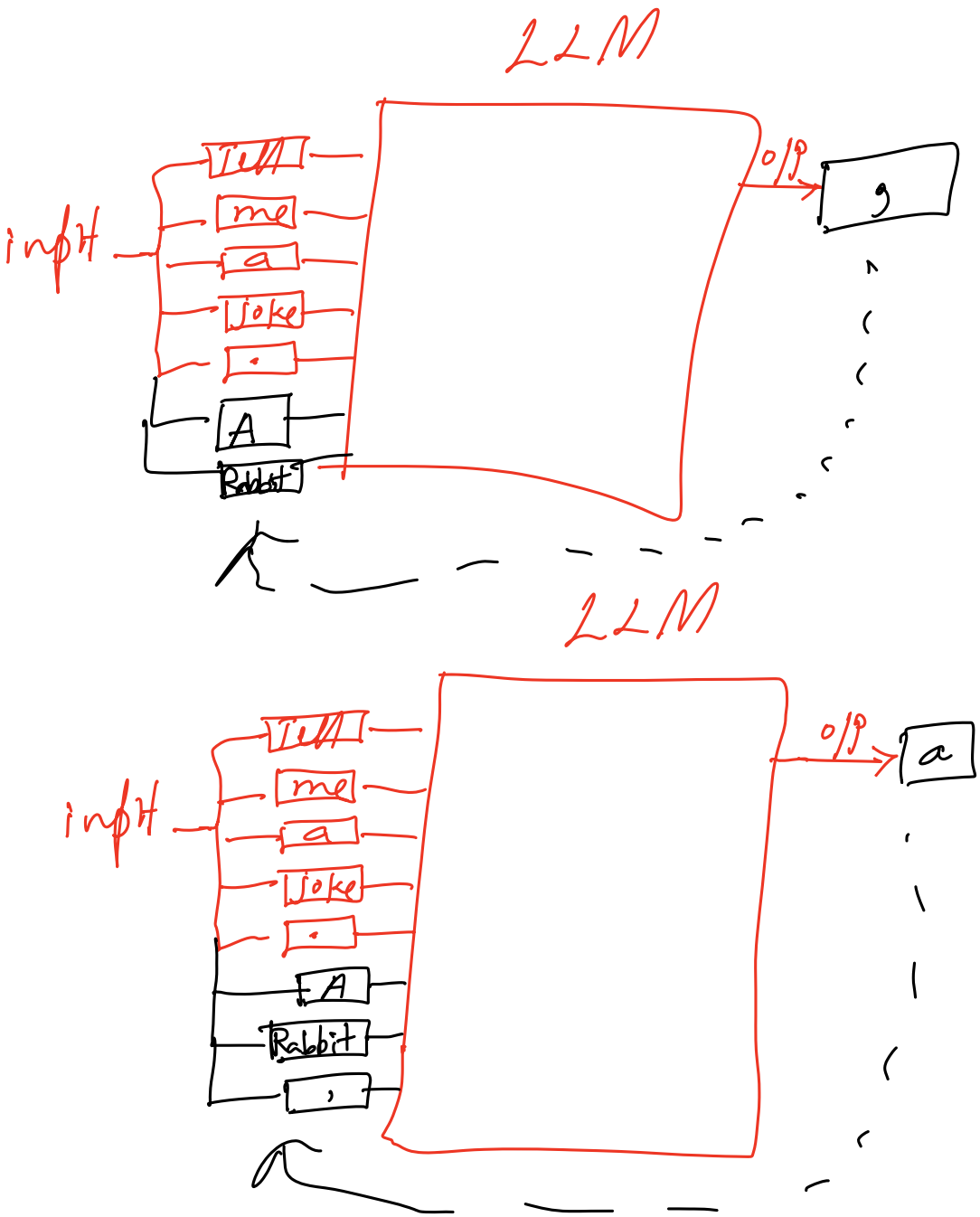
Response: A Rabbit, a lion, and a deer... <END>

LLM

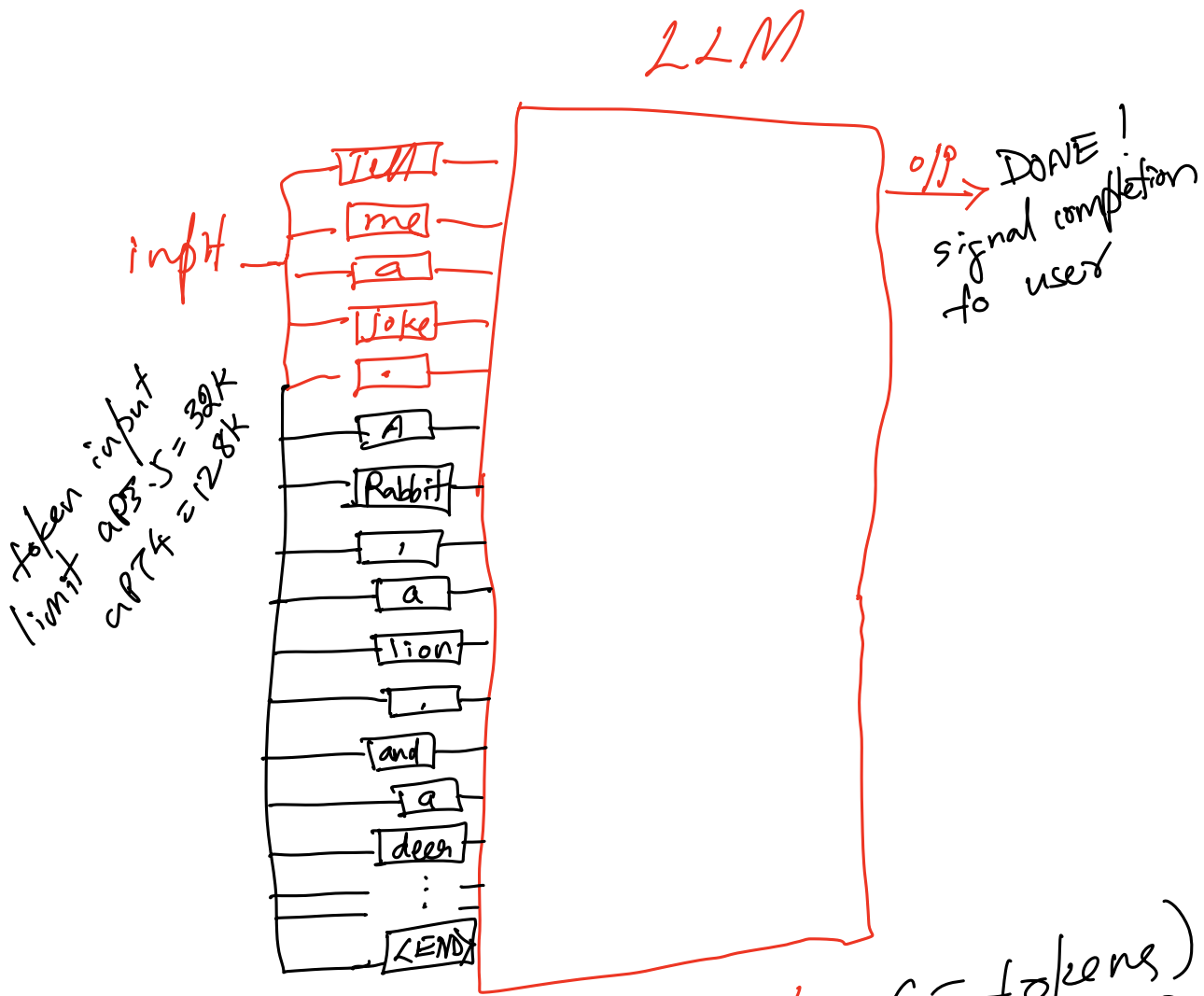
FOR A GIVEN TOKEN SEQ, PREDICT THE NEXT WORD






LLM
inpH
Tell
me
a
joke
.
A
Rabbit
o/p
g
LLM
inpH
Tell
me
a
joke
.
A
Rabbit
,
o/p
a

Finally,

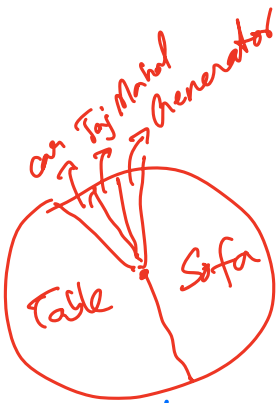


Prompt: Tell me a joke. (5 tokens)

Response: A rabbit, ... (26 tokens)

LLMs are good at generating "THE NEXT BEST WORD"

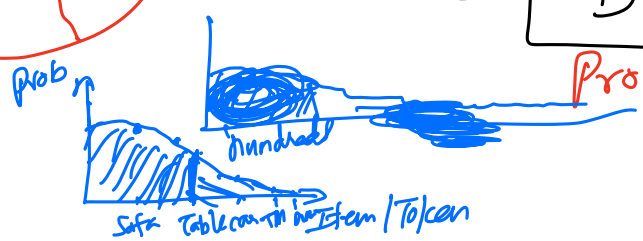
The cat jumped on the



random select

		Probability
0	Sofa	0.45
1	Table	0.40
2	car	0.10
3	Taj Mahal	0.025
4	Generator	0.025

Probability Distribution Table



User: Tell me a joke → LLM

Sys: A Rabbit, a . . . ←

User: Tell me a joke →

There was a barber . . . ←

DEMO ON PLAYGROUND

The cat is good

The cat is a liar

The cat is fat

George is a good liar

The cat is a good liar