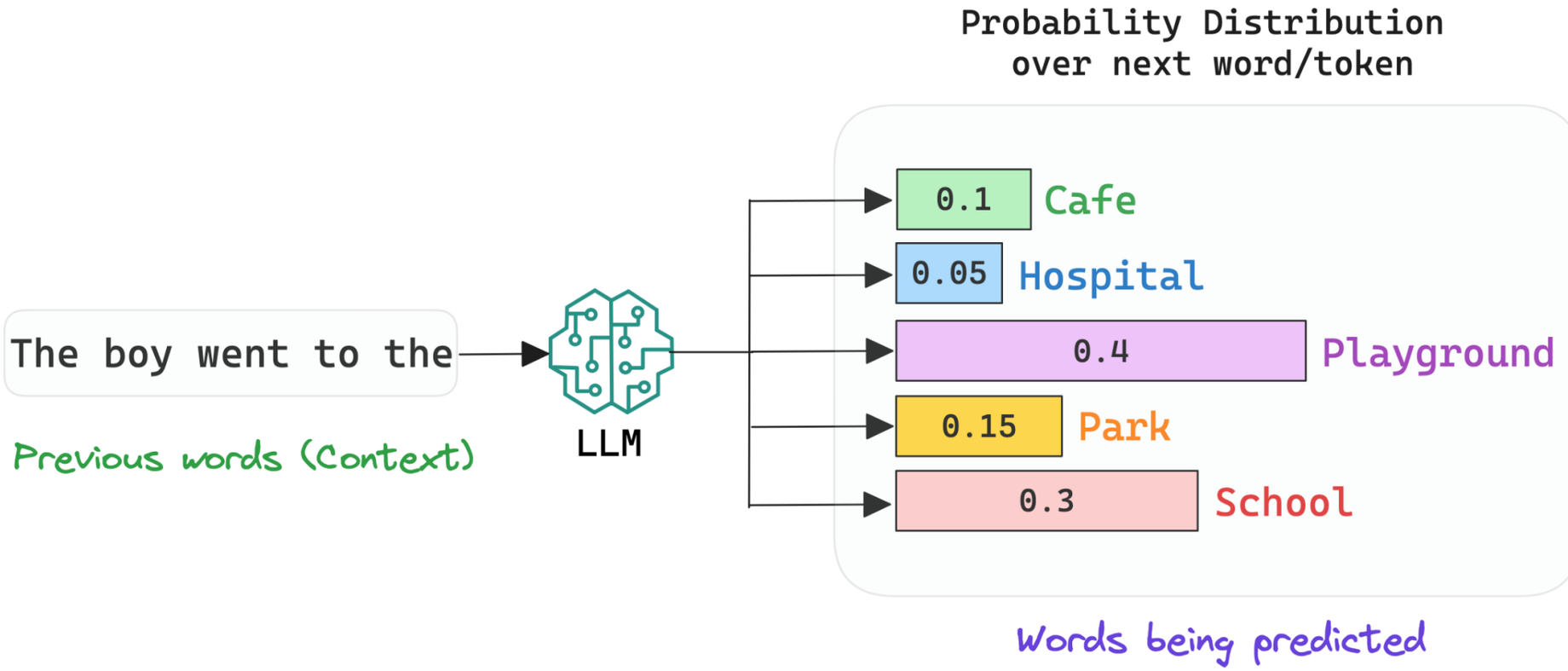
How Language models work !?



P('Playground'/'The boy went to') = 0.4

Probability that the next word is 'Playground' given that the context is 'The boy went to the'

@akshay_pachaar

Before diving into LLMs, lets understand conditional probability.

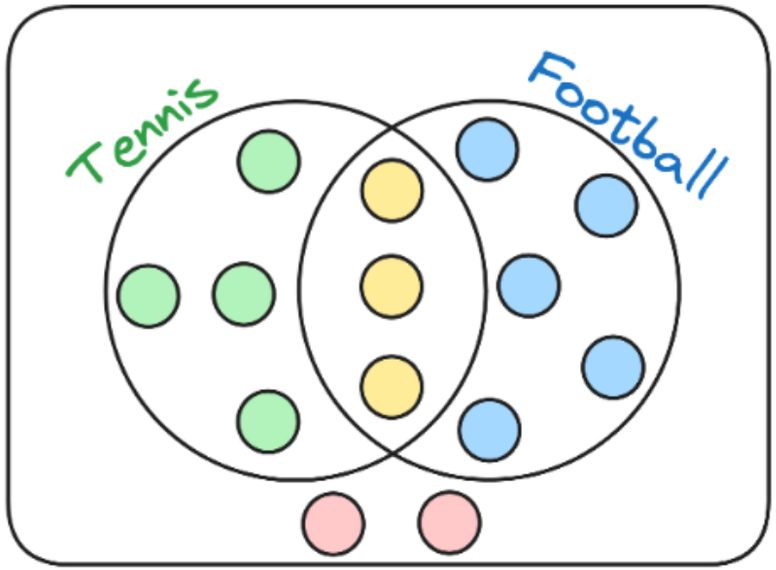
We consider a population of 14 individuals:

- Some of them like Tennis 🎾
- Some like Football ⚽
- A few like both 🎾 ⚽
- And few like none

Swipe 👉

@akshay_pachaar

A view of the Population!



	Tennis ✓	Tennis ✗
Football ✓	3	5
Football ✗	4	2

Population size = 14

A: An individual loves Tennis

B: An individual loves Football

P(A)=7/14

P(B)=8/14

@akshay_pachaar

So what is Conditional probability ⁉️

It's a measure of the probability of an event given that another event has occurred.

If the events are A and B, we denote this as P(A|B).

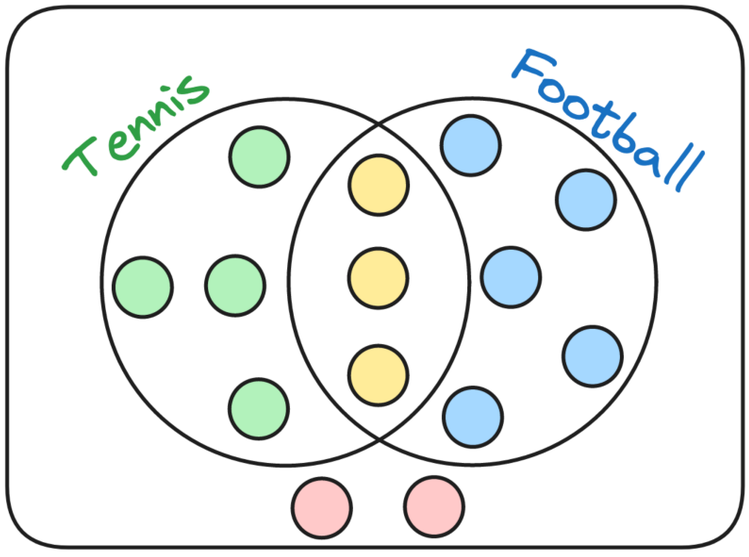
This reads as "probability of A given B"

Swipe 👉

@akshay_pachaar

Conditional Probability!

Population size = 14



	Tennis ✓	Tennis ✗
Football ✓	3	5
Football ✗	4	2

A∩B:Loves Tennis & Football

P(A∩B)=3/14

A: Loves Tennis

P(A)=7/14

B: Loves Football

P(B)=8/14

$$P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{(3/14)}{(8/14)} = 3/8$$

Probability of A given B

@akshay_pachaar

For instance, if we're predicting whether it will rain today (event A), knowing that it's cloudy (event B) might impact our prediction.

As it's more likely to rain when it's cloudy, we'd say the conditional probability P(A|B) is high.

That's conditional probability for you! 🎉

Swipe 👉

@akshay_pachaar

Now, how does this apply to LLMs like GPT-4❓

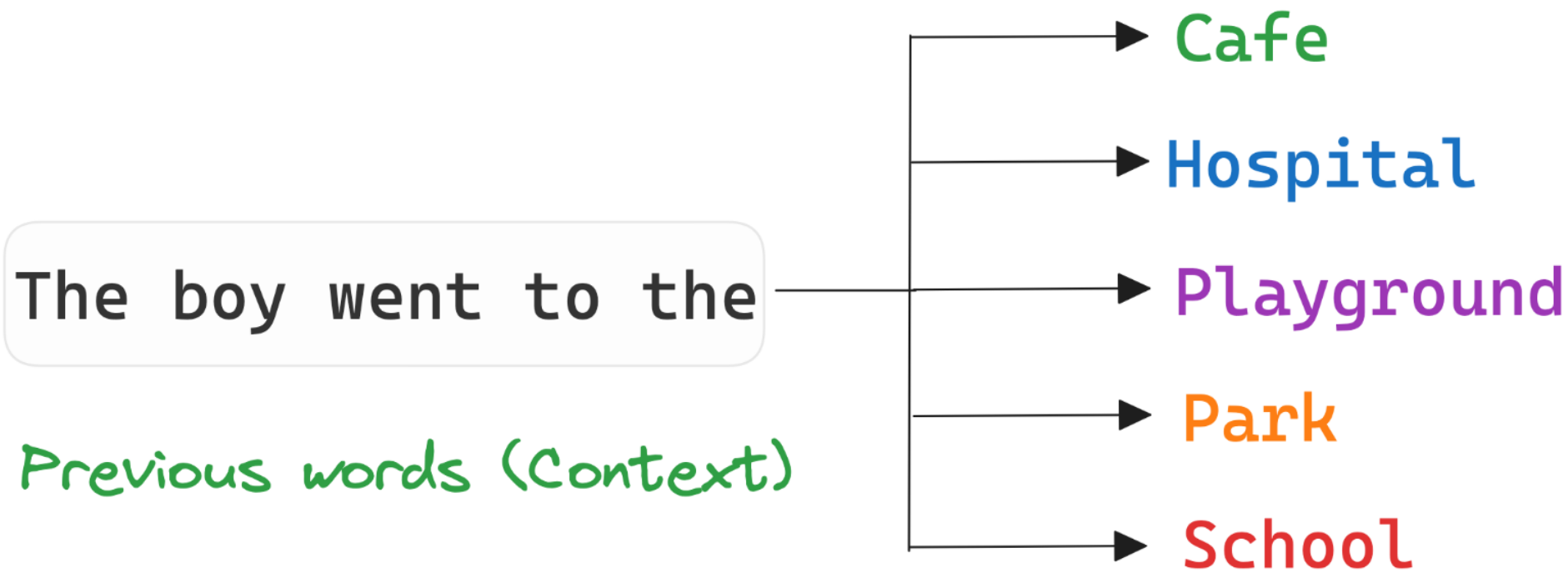
These models are tasked with predicting the next word in a sequence.

This is a question of conditional probability: given the words that have come before, what is the most likely next word?

Swipe 👉

@akshay_pachaar

The most likely next word!?



@akshay_pachaar

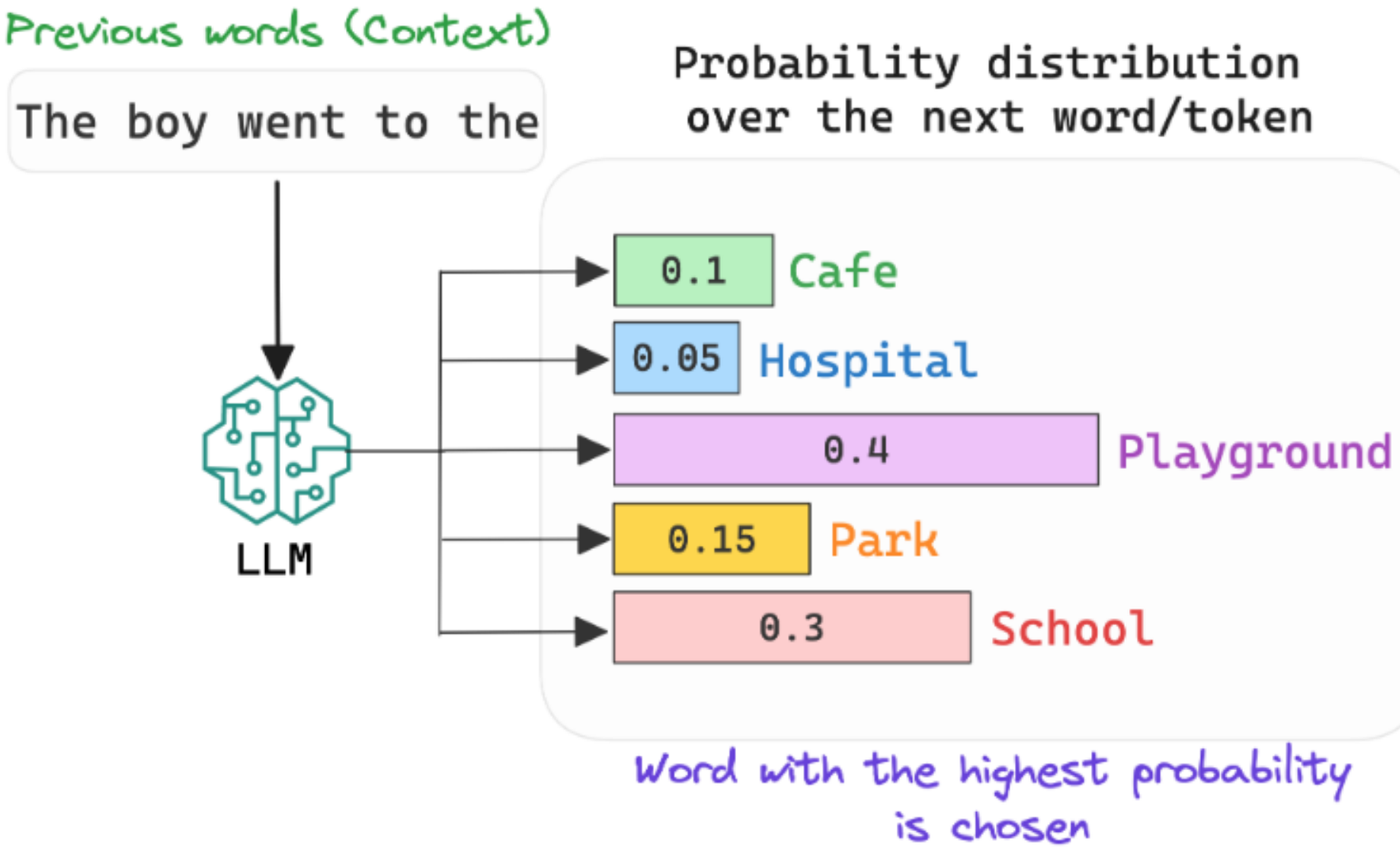
To predict the next word, the model calculates the conditional probability for each possible next word, given the previous words (context).

The word with the highest conditional prob.is chosen as the prediction.

Swipe 👉

@akshay_pachaar

Next word prediction!



@akshay_pachaar

The LLM learns a high-dimensional probability distribution over sequences of words.

And the parameters of this distribution are the trained weights!

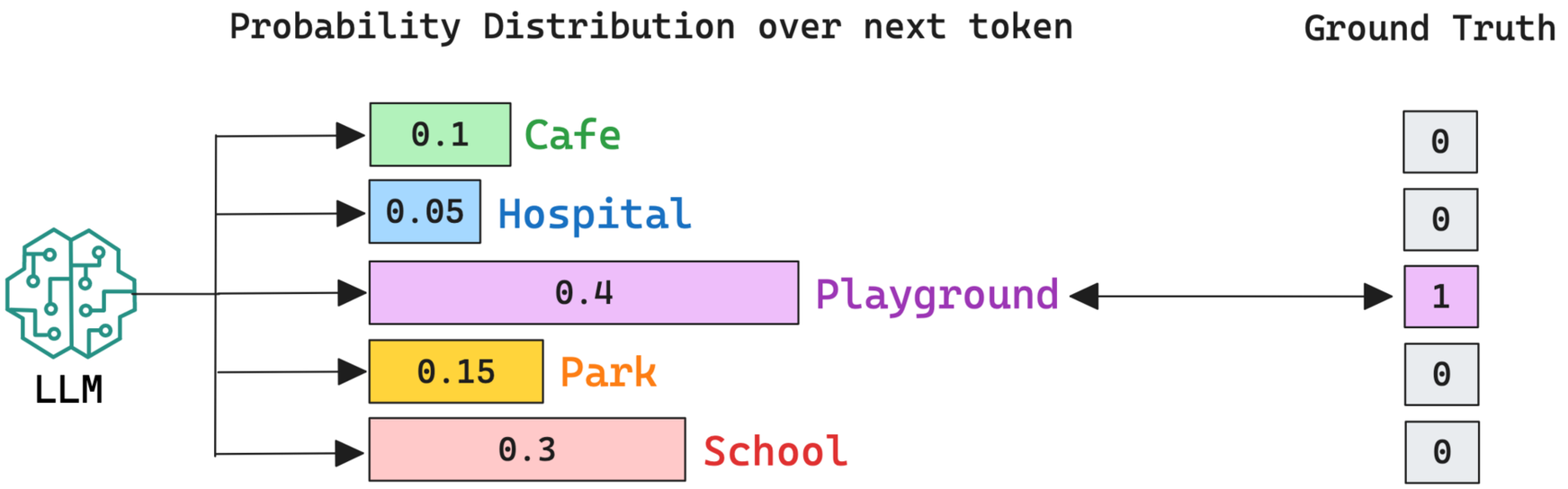
The training or rather pre-training** is supervised.

I'll talk about the different training steps next time!**

Swipe 👉

@akshay_pachaar

Loss calculation ⁉️



Loss = -log(P('Playground'/'The boy went to')) = -log(0.4)

Cross-entropy loss / Negative log-likelihood

@akshay_pachaar

Hopefully, this explanation has demystified a bit of the magic behind LLMs and the concept of conditional probability.

If you have more questions or there are other topics you want me to unpack, feel free to comment below!

Swipe 👉

@akshay_pachaar

That's a wrap!

If you interested in:

- Python 🐍
- Data Science 📈
- Machine Learning 🤖
- MLOps 🛠️
- NLP 🗣️
- Computer Vision 📽️
- LLMs 🧠

Follow me on LinkedIn ✔️
Everyday, I share tutorials on above topics!

Cheers!! 🙂

@akshay_pachaar