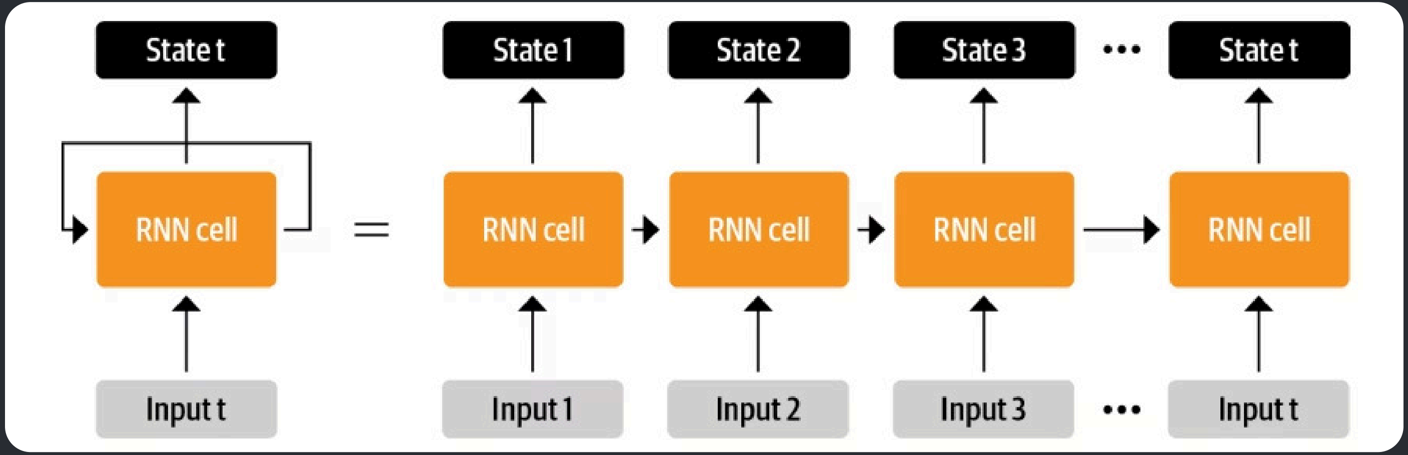
Generative AI:

An Overview

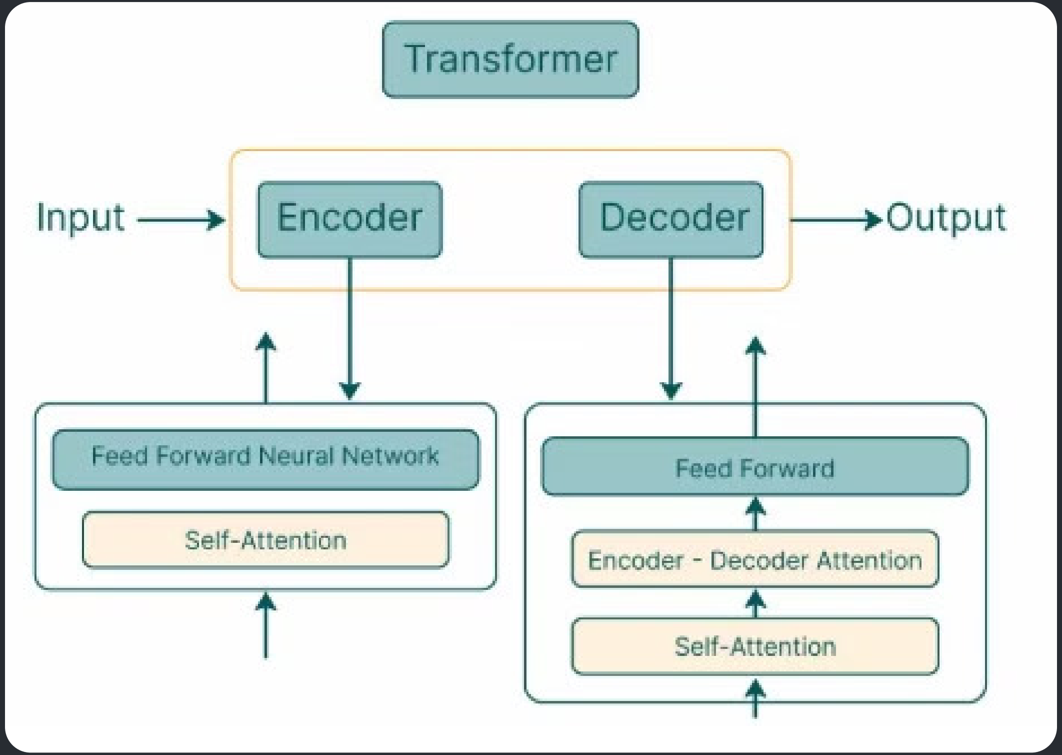
Made with GAMMA

Understanding Recurrent Neural Networks (RNNs)

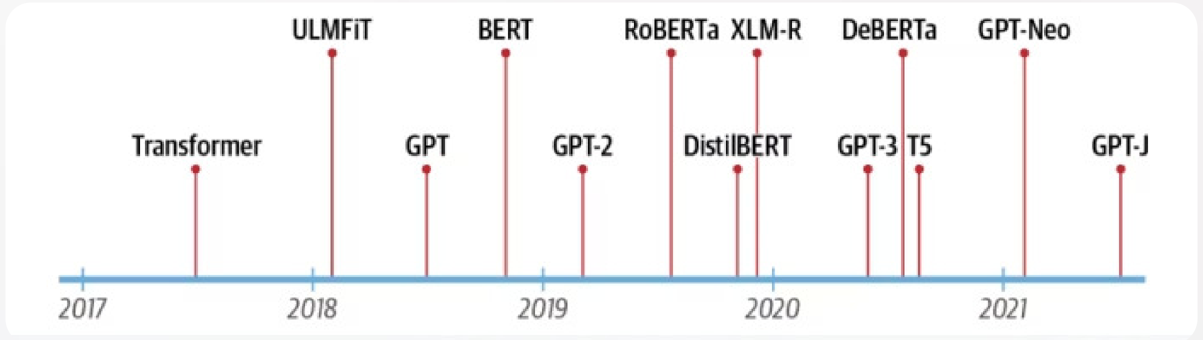


- RNNs are a type of neural network. They are designed to process sequential data.
- These architectures were widely used for NLP tasks, speech processing, and time series.
- Challenge-?

The Rise of Transformers: Self-Attention



- In 2017, researchers at Google published a paper that proposed a novel neural network architecture for sequence modeling known as Transformer.
- Outperformed recurrent neural networks (RNNs) on machine translation tasks, both in terms of translation quality and training cost.



A Timeline of Large Language Models

2022: ChatGPT

Generative Pre-trained Transformer 2.

2024: Meta's Llama 3, Claude 3, and Q2, and Mistral's Mixtral 8x7B

Larger and more powerful model.

2025: DeepSeek-R1

Multimodality: Text, Image, Video

Made with GAMMA

Diving into ChatGPT

Generative

Next word prediction

Pre-trained

LLM is pre-trained on massive amount of text

Transformer

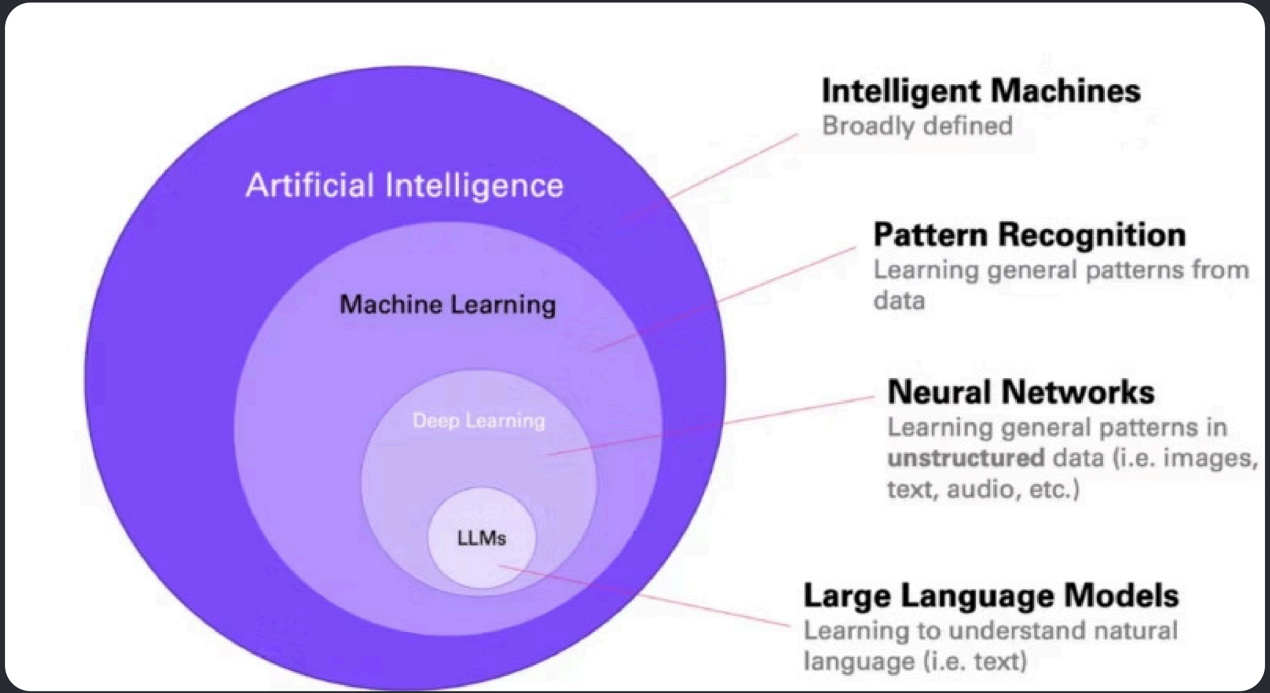
Encoder-decoder architecture

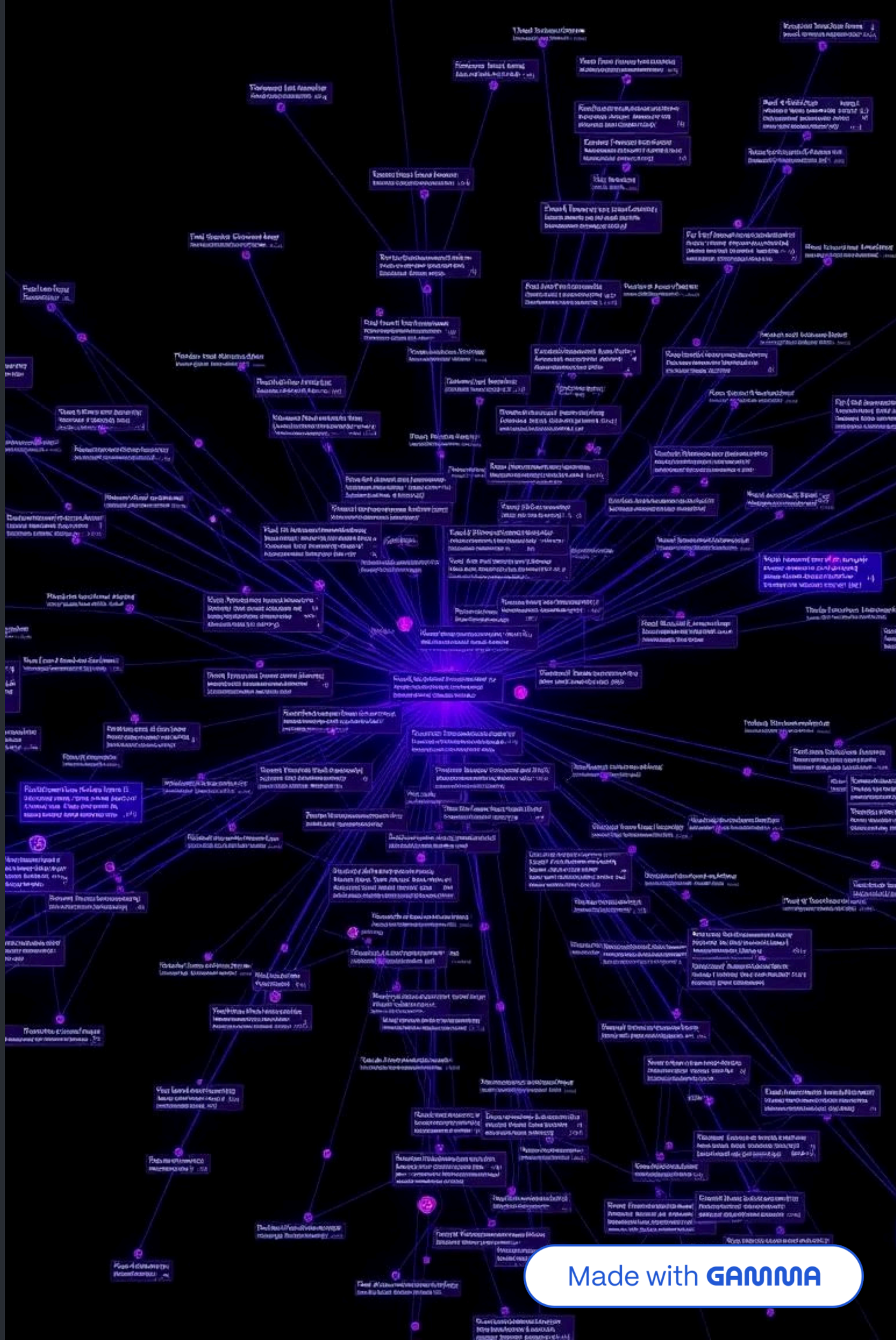
Why did ChatGPT couldn't replace Google Search?

How was ChatGPT trained?

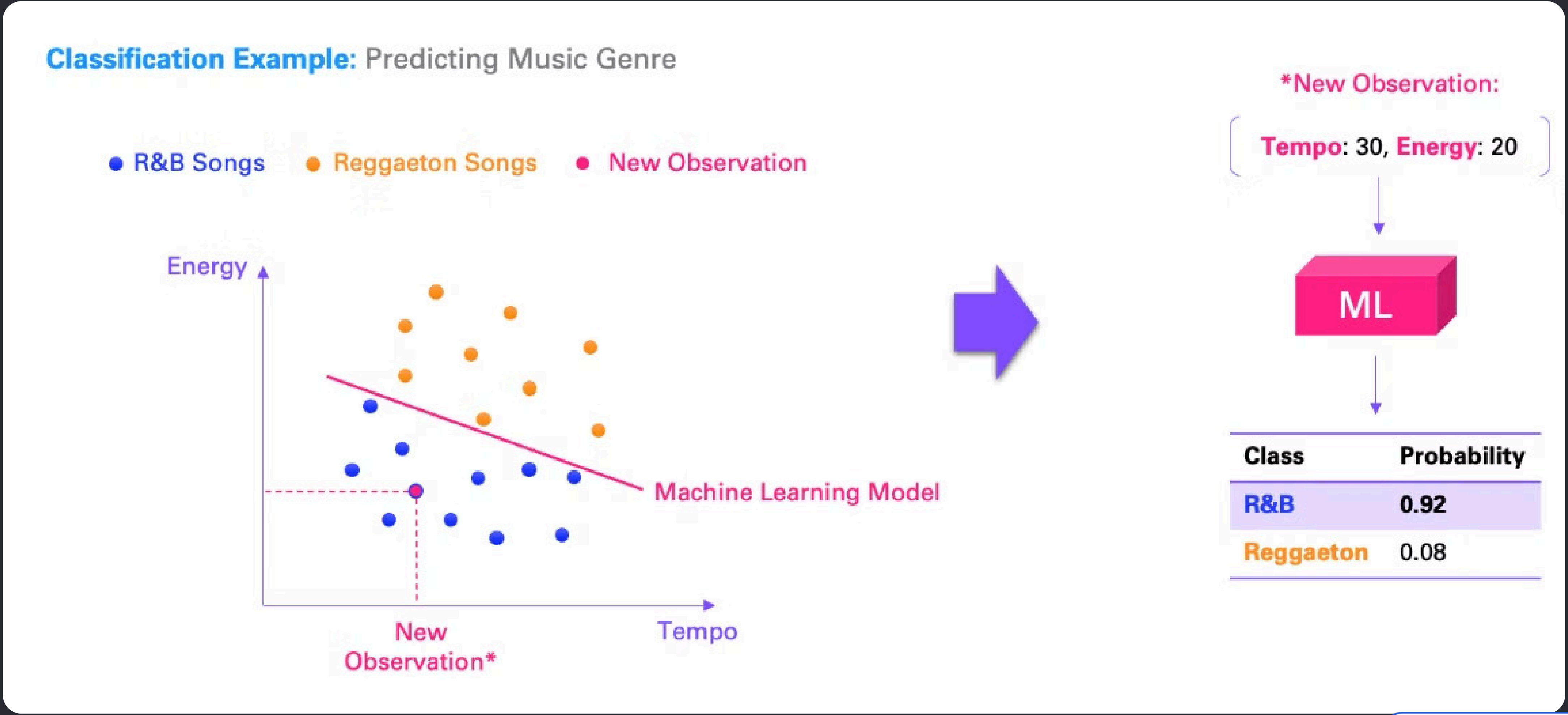
Large Language Models

What do LLMs essentially do?


Artificial Intelligence
Machine Learning
Deep Learning
LLMs
Intelligent Machines
Broadly defined
Pattern Recognition
Learning general patterns from data
Neural Networks
Learning general patterns in **unstructured** data (i.e. images, text, audio, etc.)
Large Language Models
Learning to understand natural language (i.e. text)



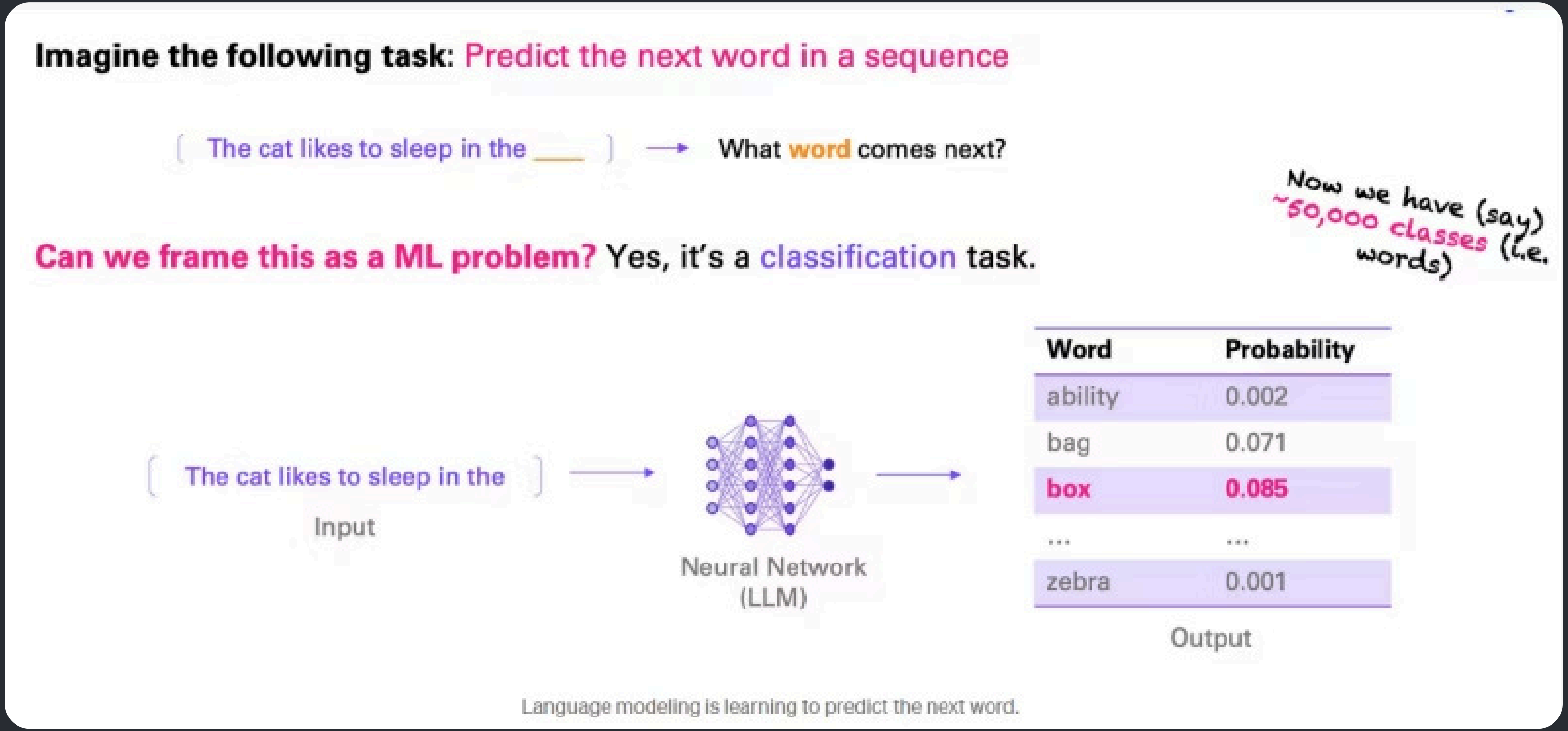
LLMs as Machine Learning Task?



Classification Example: Predicting Music Genre

Class	Probability
R&B	0.92
Reggaeton	0.08

LLMs as Deep Learning Task?



Imagine the following task: Predict the next word in a sequence

[The cat likes to sleep in the ____] → What **word** comes next?

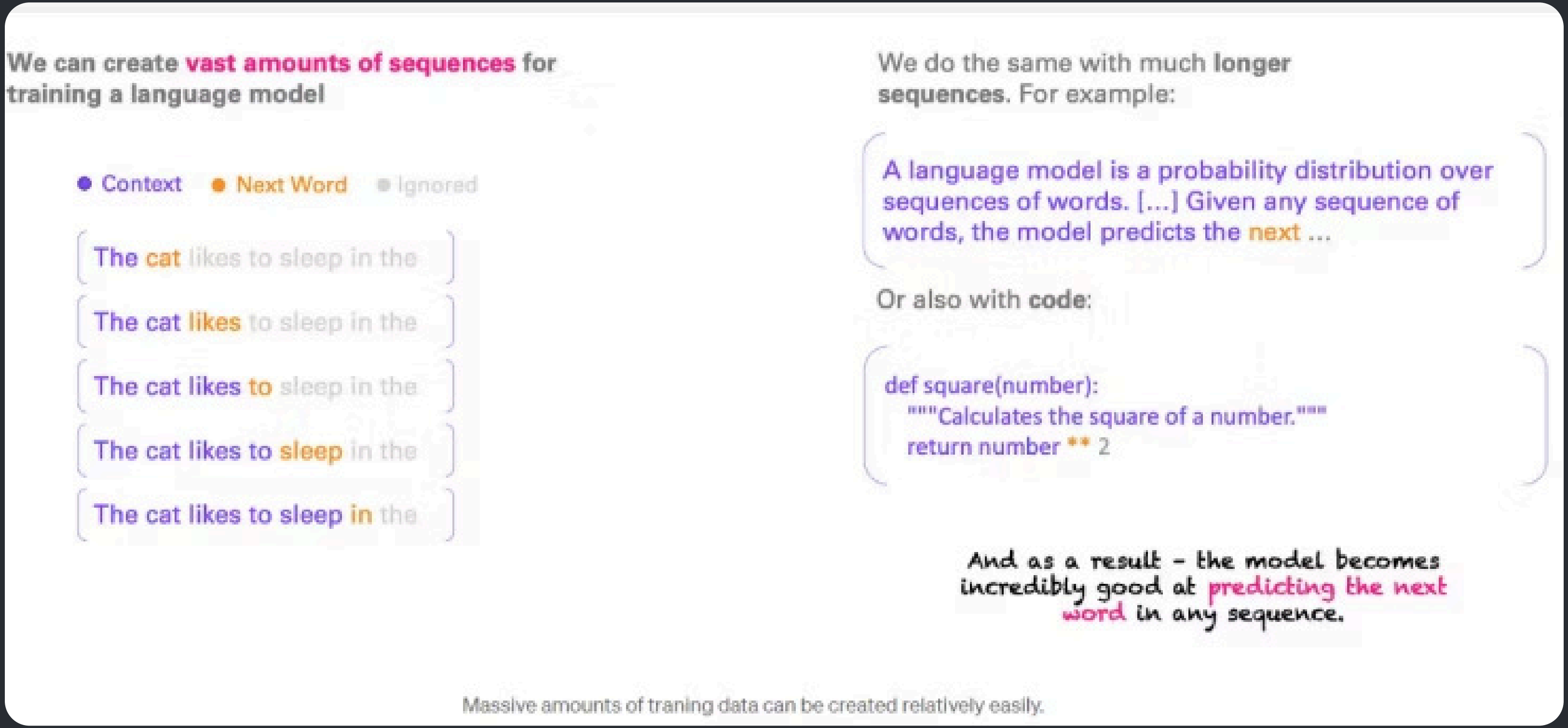
Now we have (say) ~50,000 classes (i.e. words)

Can we frame this as a ML problem? Yes, it's a classification task.

Word	Probability
ability	0.002
bag	0.071
box	**0.085**
...	...
zebra	0.001

Language modeling is learning to predict the next word.

Training Data for LLMs



We can create vast amounts of sequences for training a language model

● Context ● Next Word ● Ignored

The cat likes to sleep in the

The cat likes to sleep in the

The cat likes to sleep in the

The cat likes to sleep in the

The cat likes to sleep in the

We do the same with much longer sequences. For example:

A language model is a probability distribution over sequences of words. [...] Given any sequence of words, the model predicts the next ...

Or also with code:

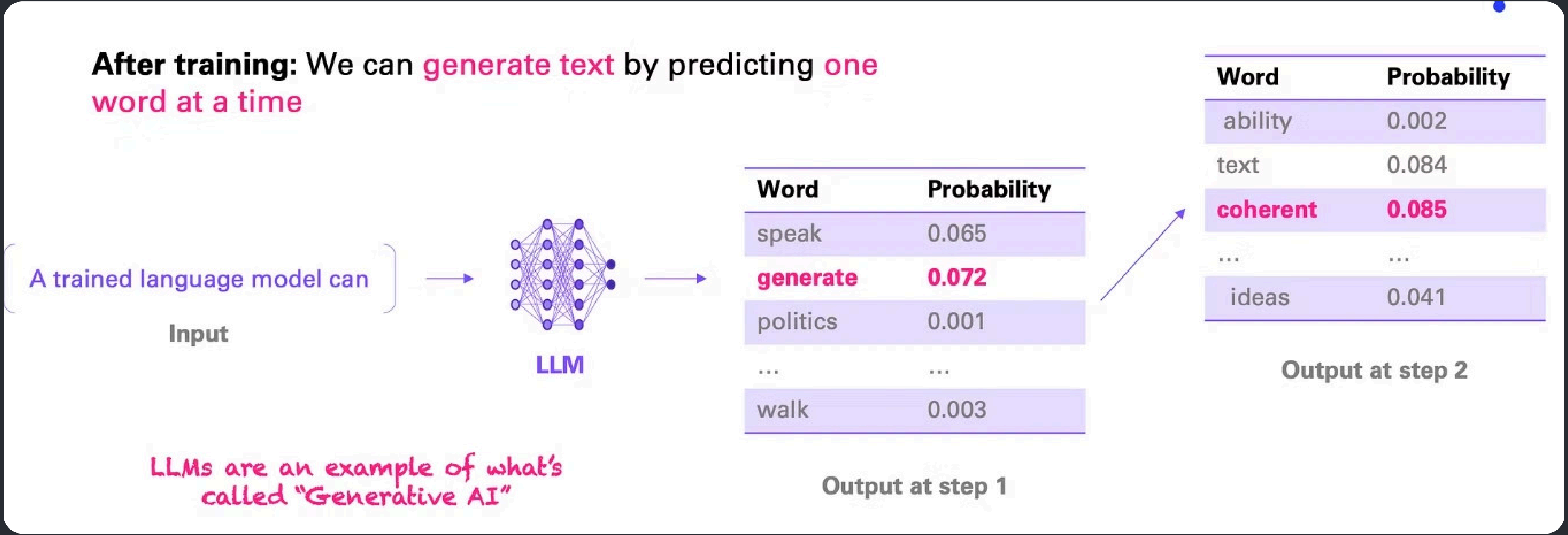
```
def square(number):
    """Calculates the square of a number."""
    return number ** 2
```

And as a result - the model becomes incredibly good at predicting the next word in any sequence.

Massive amounts of traning data can be created relatively easily.

Made with GAMMA

Next word Generation



After training: We can generate text by predicting one word at a time

Word	Probability
speak	0.065
generate	**0.072**
politics	0.001
...	...
walk	0.003

Output at step 1

Word	Probability
ability	0.002
text	0.084
coherent	**0.085**
...	...
ideas	0.041

Output at step 2

LLMs are an example of what's called "Generative AI"

Phases of LLM Training

Pre-training

- Massive amount of text data from internet - books, research papers, websites
- Model learns to predict the next word

Instruction fine tuning

- Curating Q n A dataset to train the model to answer questions or instructions
- Model learns to become a helpful assistant

Reinforcement Learning from Human Feedback (RLHF)

- Align the output closer to human like responses
- Responses are updated considering human feedback and preference.

Limitation of LLMs

1. Hallucination
2. Mathematical Problem solving
3. Context window
4. Cost

How to make LLMs respond better?

Zero-Shot

- Give some instructions to solve a task.

Few-Shot

- Give some examples of how to solve a task.

Chain-of-Thought(CoT)

- For complex tasks- prompt an LLM to “think step by step”

Latest LLMs & Frameworks

LLMs

Mistral

Mixtral

Llama

Gemini

DeepSeek

Frameworks

Together AI- https://www.together.ai/

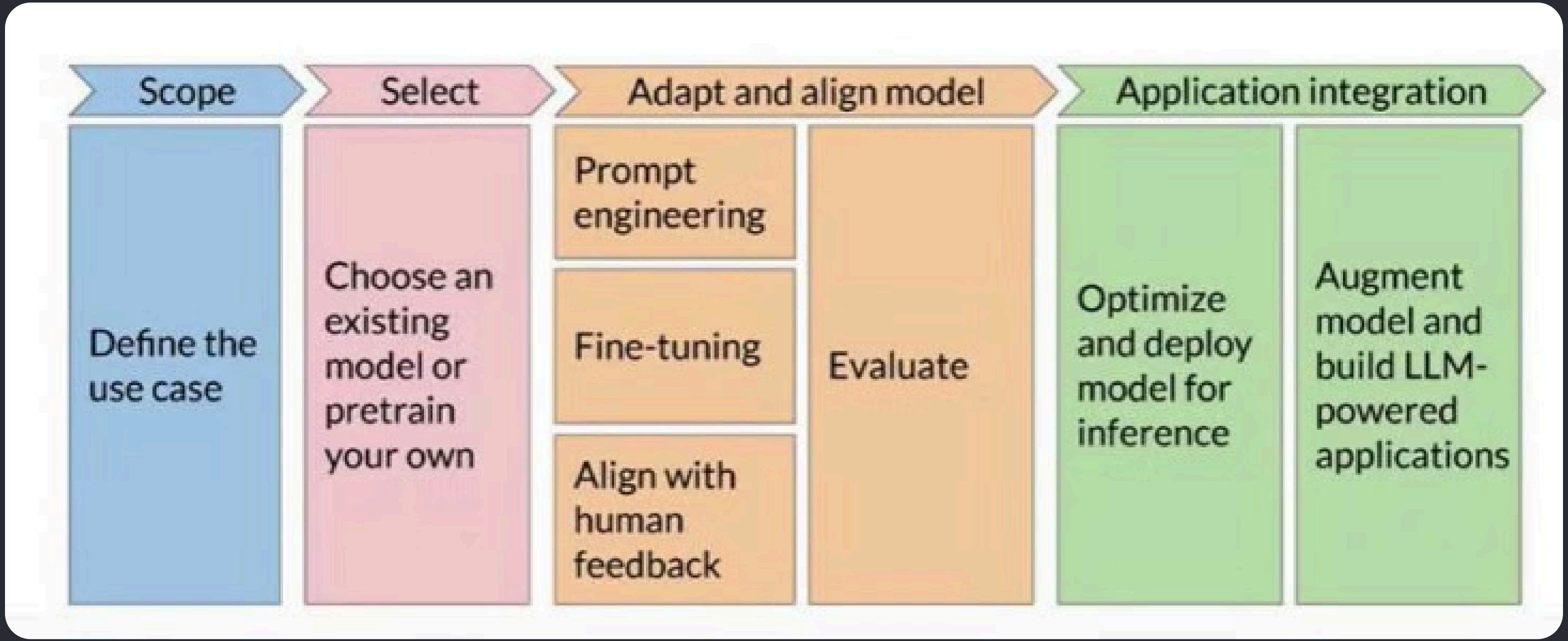
Groq- https://groq.com/

Replicate- https://replicate.com/

LiteLLM - https://www.litellm.ai/

Hugging Face- https://huggingface.co/

Generative AI Project Lifecycle



Scope
Select
Adapt and align model
Application integration
Define the use case
Choose an existing model or pretrain your own
Prompt engineering
Fine-tuning
Align with human feedback
Evaluate
Optimize and deploy model for inference
Augment model and build LLM-powered applications