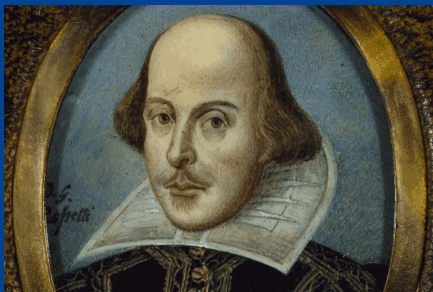




## Lesson 11: Introduction to RNN, LLM and GPT

CARSTEN EIE FRIGAARD

AUTUMN 2024

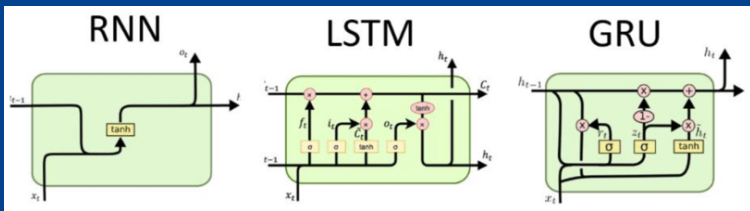


# Agenda

- ▶ Search Quest and Cake...
- ▶ Intro to text-processing in ML
  - ▶ RNNs,
  - ▶ LLMs,
  - ▶ GPTs
- ▶ ..and hands-on exercise nanoGPT.

# INTRO TO ML TEXT PROCESSING

---



# Recurrent Neural Networks (RNN)

network against the time axis, as shown in Figure 15-1 (right). This is called *unrolling the network through time* (it's the same recurrent neuron represented once per time step).

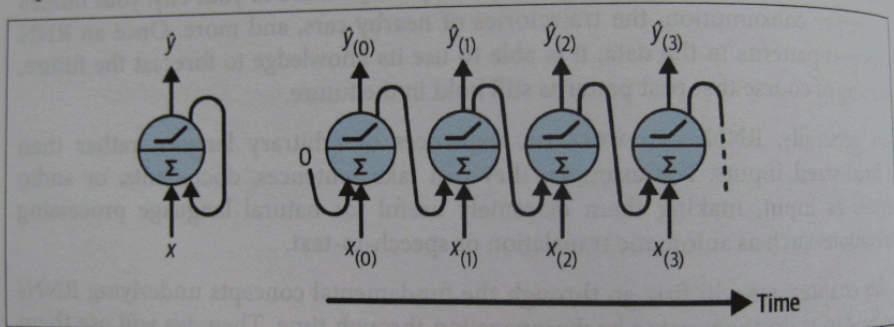


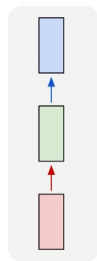
Figure 15-1. A recurrent neuron (left) unrolled through time (right)

You can easily create a layer of  $n$  neurons. At each time step  $t$ , every neuron receives both the input vector  $x_{(t)}$  and the output vector from the previous time step  $\hat{y}_{(t-1)}$ , as shown in Figure 15-2. With the inputs and outputs are now vectors (when there was just a single neuron).

# RNN

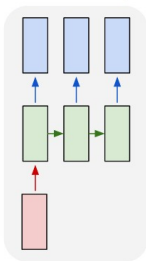
## Recurrent Neural Networks: Sequences and Transformation...

one to one



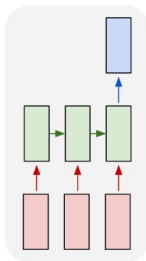
*image classification*

one to many



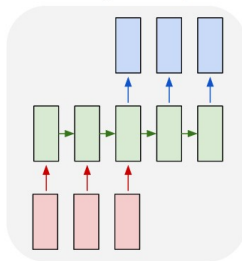
*image captioning*

many to one



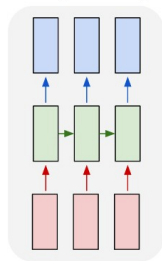
*sentiment analysis*

many to many



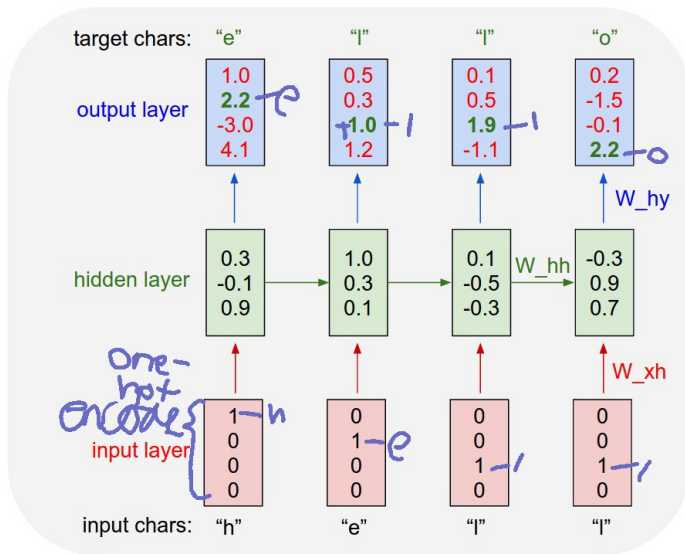
*machine translation*

many to many



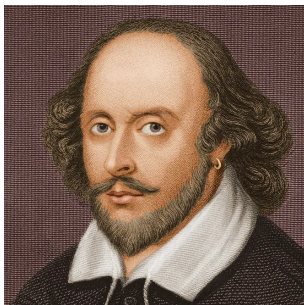
# RNN

## Transformer: input-output encoding



# DATASET: HCA or Shakespeare (or DR news)

Text data and transformers..



*"Der kom en soldat marcherende hen ad landevejen [...]"*

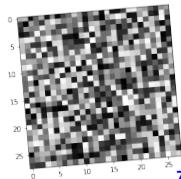
"Der kom en sol"      => 'd'

"er kom en sold"      => 'a'

"r kom en solda"      => 't'

" kom en soldat"      => ' '

Random text and image: "vim5 Rrd vjmt8vt"



# LLM

---

Model	Model architecture	Training data	Model weights	Checkpoints	Compute-optimal training	License
OpenAI GPT-4	Closed	Closed	No	No	Unknown	Not available
Deepmind Chinchilla	Open	Closed	No	No	Yes	Not available
Meta OPT	Open	Open	Researchers Only	Yes	No	Non-commercial
Pythia	Open	Open	Open	Yes	No	Apache 2.0
Cerebras-GPT	Open	Open	Open	Yes	Yes	Apache 2.0

[[https://jacar.es/wp-content/uploads/2023/03/cerebras\\_gpt\\_models.png](https://jacar.es/wp-content/uploads/2023/03/cerebras_gpt_models.png)]





# Large language model

[45 languages](#) ▾[Article](#) [Talk](#)[Read](#) [Edit](#) [View history](#) [Tools](#) ▾

From Wikipedia, the free encyclopedia

*Not to be confused with [Logic learning machine](#).*

A **large language model (LLM)** is a type of computational [model](#) designed for [natural language processing](#) tasks such as language [generation](#). As [language models](#), LLMs acquire these abilities by [learning statistical relationships](#) from vast amounts of text during a [self-supervised](#) and [semi-supervised](#) training process.<sup>[1]</sup>

The largest and most capable LLMs are [artificial neural networks](#) built with a decoder-only [transformer-based architecture](#), enabling efficient processing and generation of large-scale text data. Modern models can be [fine-tuned](#) for specific tasks, or be guided by [prompt engineering](#).<sup>[2]</sup> These models acquire [predictive power](#) regarding [syntax](#), [semantics](#), and [ontologies](#)<sup>[3]</sup> inherent in human language

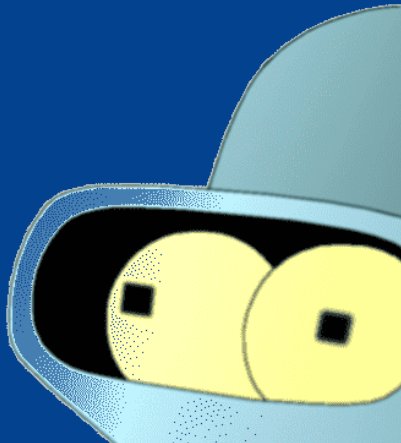
Part of a series on

## Machine learning and data mining

[Paradigms](#) [\[show\]](#)[Problems](#) [\[show\]](#)[Supervised learning](#) [\[show\]](#)  
([classification](#) • [regression](#))[Clustering](#) [\[show\]](#)[Dimensionality reduction](#) [\[show\]](#)[Structured prediction](#) [\[show\]](#)[Anomaly detection](#) [\[show\]](#)[Artificial neural network](#) [\[hide\]](#)[Autoencoder](#) • [Deep learning](#) •

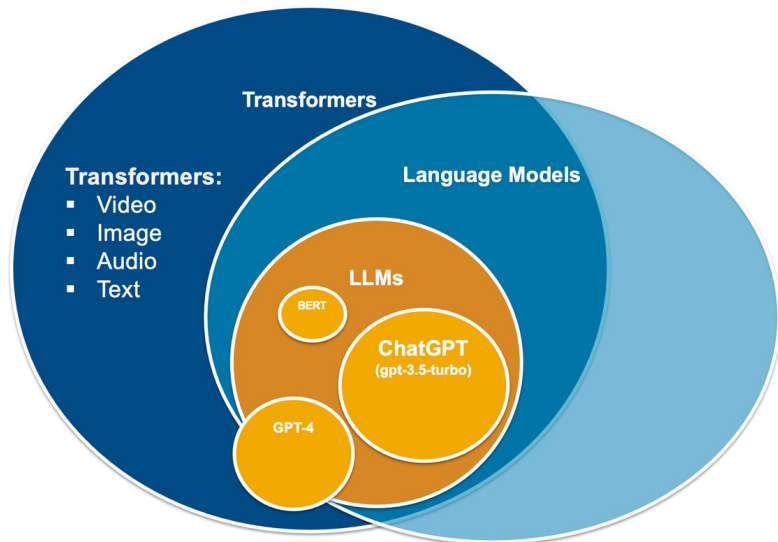
GPT

---



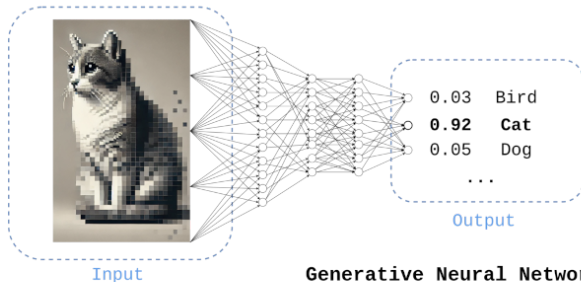
# GPT

## Generative Pre-trained Transformer

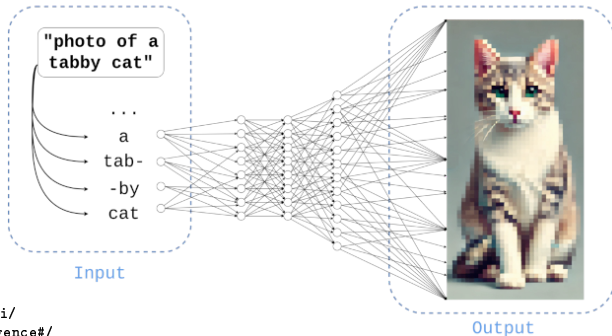


# Generative?

## Discriminative Neural Network



## Generative Neural Network

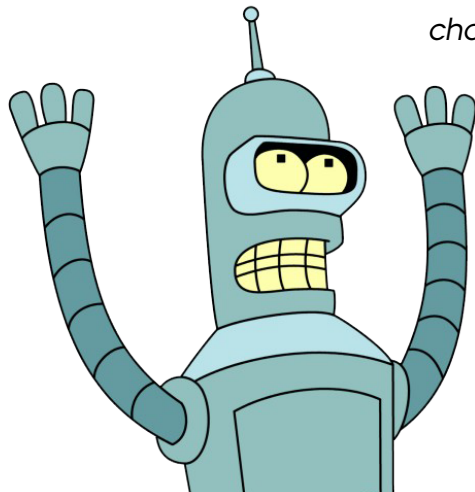


# Pre-trained?

*..is pre-trained just trained?*

*..but a Transformer alright!*

*chat-T, chat-GT, and GAI?*



# GPT HANDS-ON EXERCISE

---

available GPT implementations



~~miniGPT~~ nanoGPT



UNUSED SLIDES..

# Facts and Reflections

p7: "Temperatures are between 0 and 1. Lower temperatures inject less randomness; with a temperature of 0, ChatGPT should always give you the same response to the same prompt. If you set the temperature to 1"

p7: "For ChatGPT, the total length of the prompt and the response currently must be under 4096 tokens,"

p8: on the net "Estimates of the percentage of false statements are typically around 30

p9: "The training data for ChatGPT and GPT-4 ends in September 2021"

p11: "You will have to edit it and, while some have suggested that ChatGPT might provide a good rough draft, turning poor prose into good prose can be more difficult than writing the first draft yourself.

**NOTE:** "What Are ChatGPT and Its Friends?", Mike Loukides, O'Reilly, 978-1-098-15259-8



# Training Cost and Hardware

NVIDIA A100  
Tensor Core GPU



Once you know the parameter count, token count, and how many GPUs you have, it's easy to calculate the theoretical training costs for many popular models. For example, we will use Nvidia A100s, using \$1.5 per hour per GPU. "FLOPS utilization" will increase from 40% to 60% with larger models, explained [here](#), but generally, there isn't much room to go higher with current systems.

## State-Of-The-Art Training Costs

Model	Optimal LLM Training Cost		
	Size (# Parameters)	Tokens	GPU
MosaicML GPT-30B	30 Billion	610 Billion	A100 \$
Google LaMDA	137 Billion	168 Billion	A100 \$
Yandex YaLM	100 Billion	300 Billion	A100 \$
Tsinghua University Zhipu AI GLM	130 Billion	400 Billion	A100 \$
Open AI GPT-3	175 Billion	300 Billion	A100 \$
AI21 Jurassic	178 Billion	300 Billion	A100 \$
Bloom	176 Billion	300 Billion	A100 \$
DeepMind Gopher	280 Billion	366 Billion	A100 \$
DeepMind Chinchilla	70 Billion	300 Billion	A100 \$
MosaicML GPT-70B	70 Billion	1,400 Billion	A100 \$
Nvidia Microsoft MT-NLG	530 Billion	1,400 Billion	A100 \$
Google PaLM	540 Billion	270 Billion	A100 \$
		780 Billion	A100 \$

How much does ChatGPT cost? 2-12 million per training for large models/

TECHGOING

Threza Gabriel | 18/02/2023

How much does ChatGPT cost? \$2-12 million per training for large models



SHARE Facebook Twitter LinkedIn Reddit

ChatGPT took the world by storm, technology giants have entered the game, and generative AI behind its large model-based artificial intelligence has become the direction of industry investment.



## Latest

Blackview Tab released: 11-inch, priced at \$399  
21/03/2023

Realme 1009,127 P...  
21/03/2023

Xiaomi off...  
Redmi Ne...  
conference  
March 28  
21/03/2023

Hozon's shipping...  
21/03/2023

# Training Cost and Hardware

The Company & its Products ▾ | Bloomberg Terminal Demo Request

Bloomberg

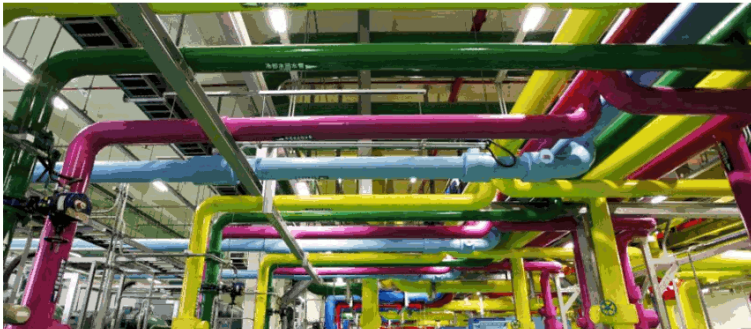
Subscribe



Green | New Energy

## Artificial Intelligence Is Booming—So Is Its Carbon Footprint

Greater transparency on emissions could also bring more scrutiny



[https://www.bloomberg.com/news/articles/2023-03-09/  
how-much-energy-do-ai-and-chatgpt-use-no-one-knows-for-sure?leadSource=verify%20wall](https://www.bloomberg.com/news/articles/2023-03-09/how-much-energy-do-ai-and-chatgpt-use-no-one-knows-for-sure?leadSource=verify%20wall)