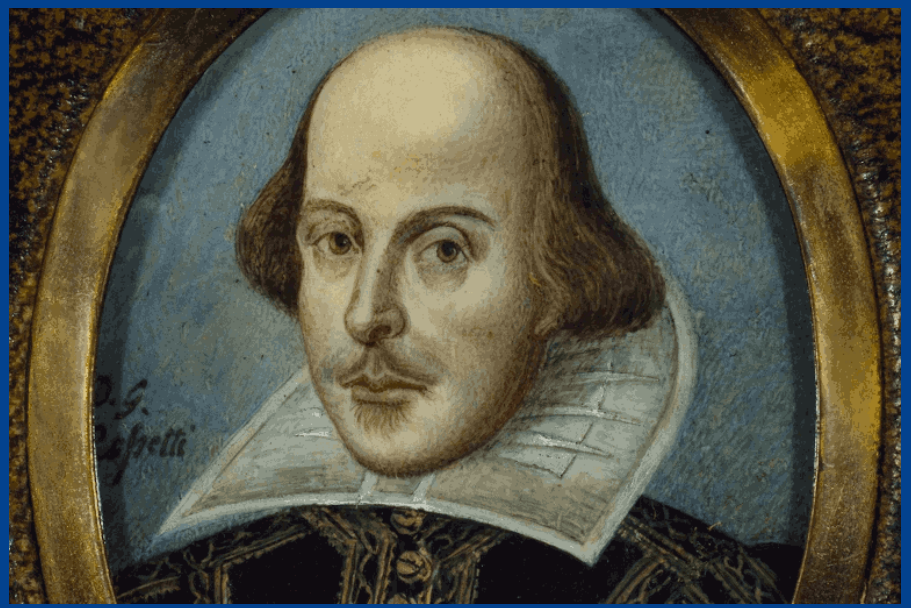
AARHUS UNIVERSITY

Lesson 11: Introduction to RNN, LLM and GPT

CARSTEN EIE FRIGAARD

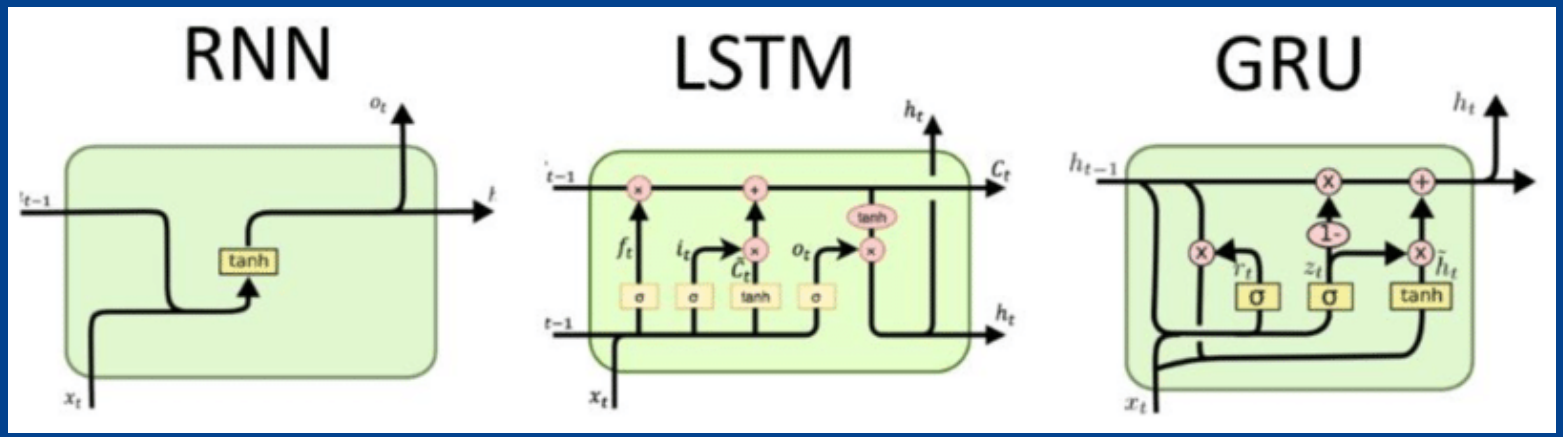
AUTUMN 2024



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).

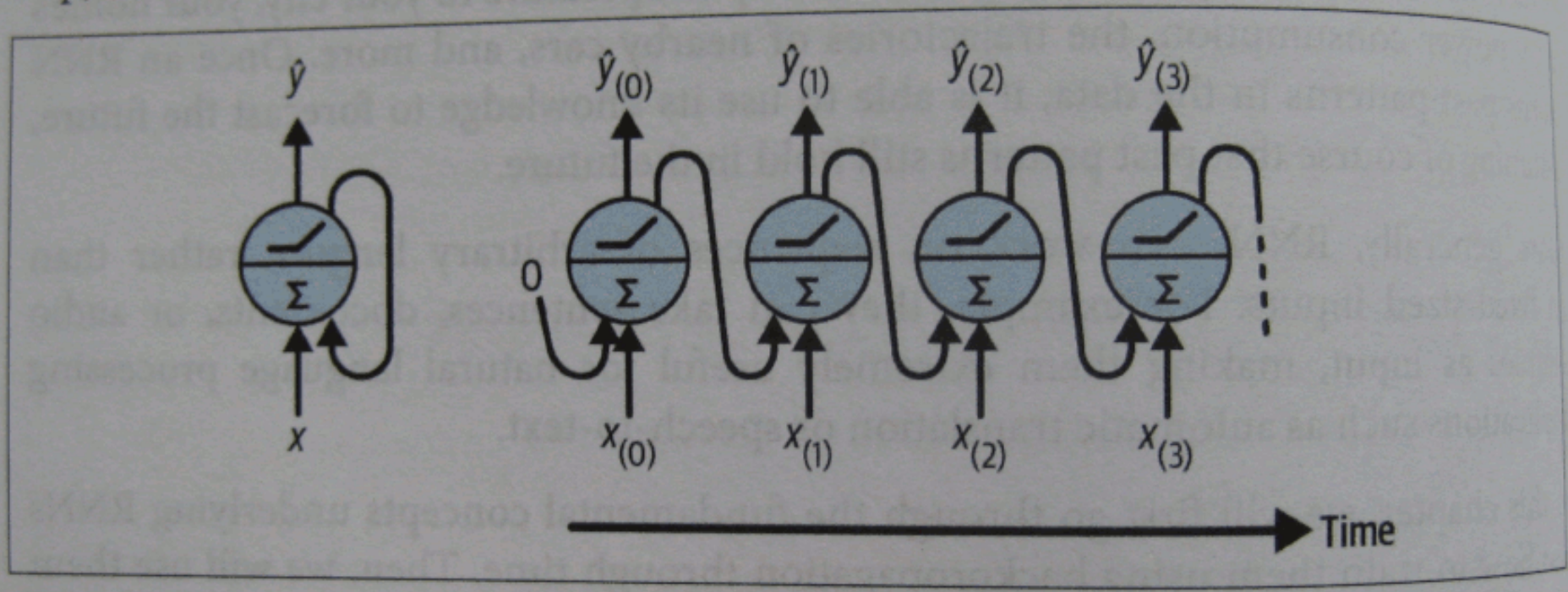


Figure 15-1. A recurrent neuron (left) unrolled through time (right)

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

RNN

Recurrent Neural Networks: Sequences and Transformation...

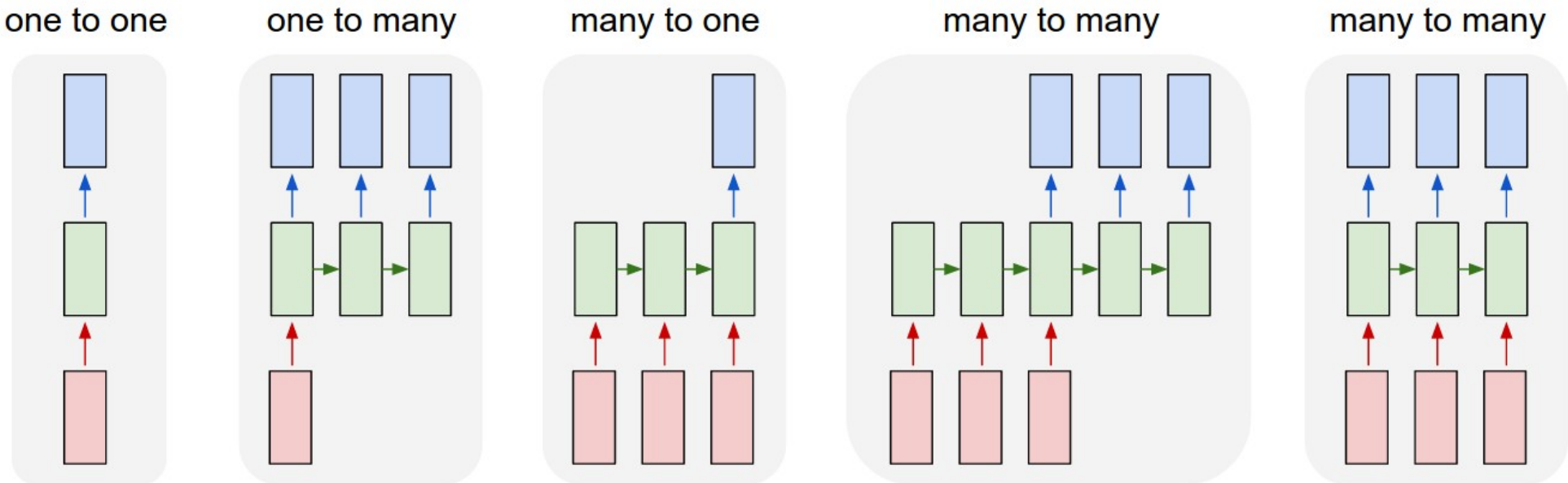


image classification

image captioning

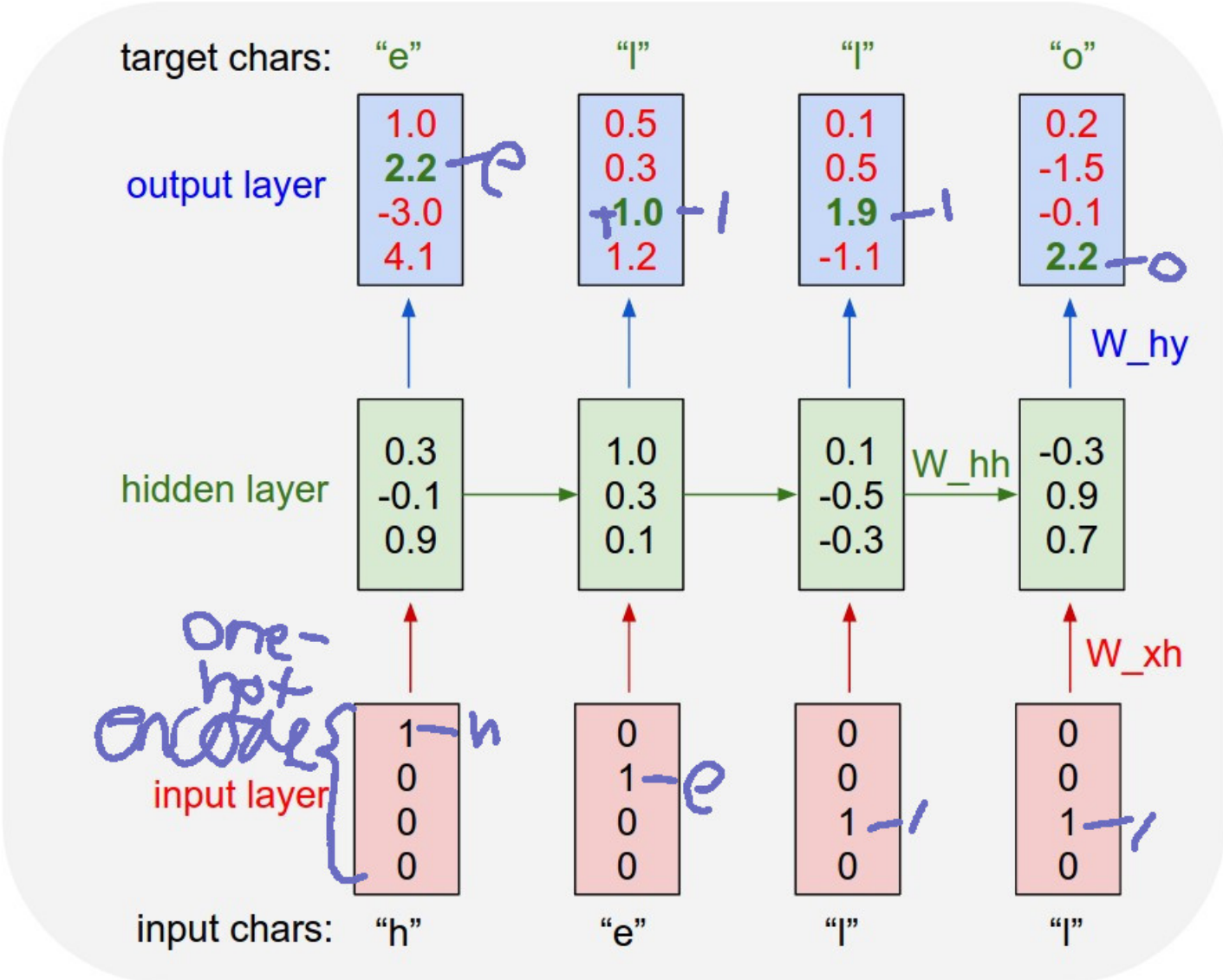
sentiment analysis

machine translation

[http://karpathy.github.io/2015/05/21/rnn-effectiveness/]

RNN

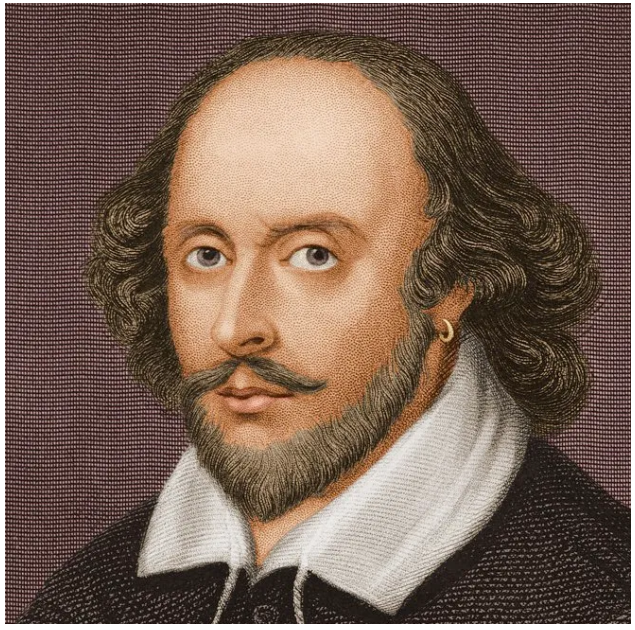
Transformer: input-output encoding



[http://karpathy.github.io/2015/05/21/rnn-effectiveness/]

DATASET: HCA or Shakespeare (or DR news)

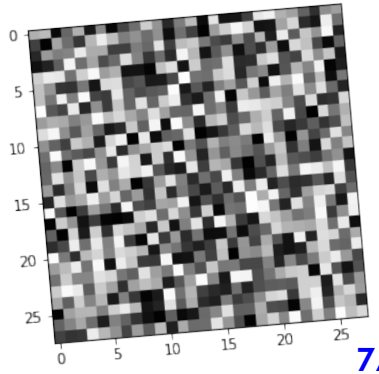
Text data and transformers..



"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]

LLM

https://en.wikipedia.org/wiki/Large_language_model

WIKIPEDIA
The Free Encyclopedia

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.

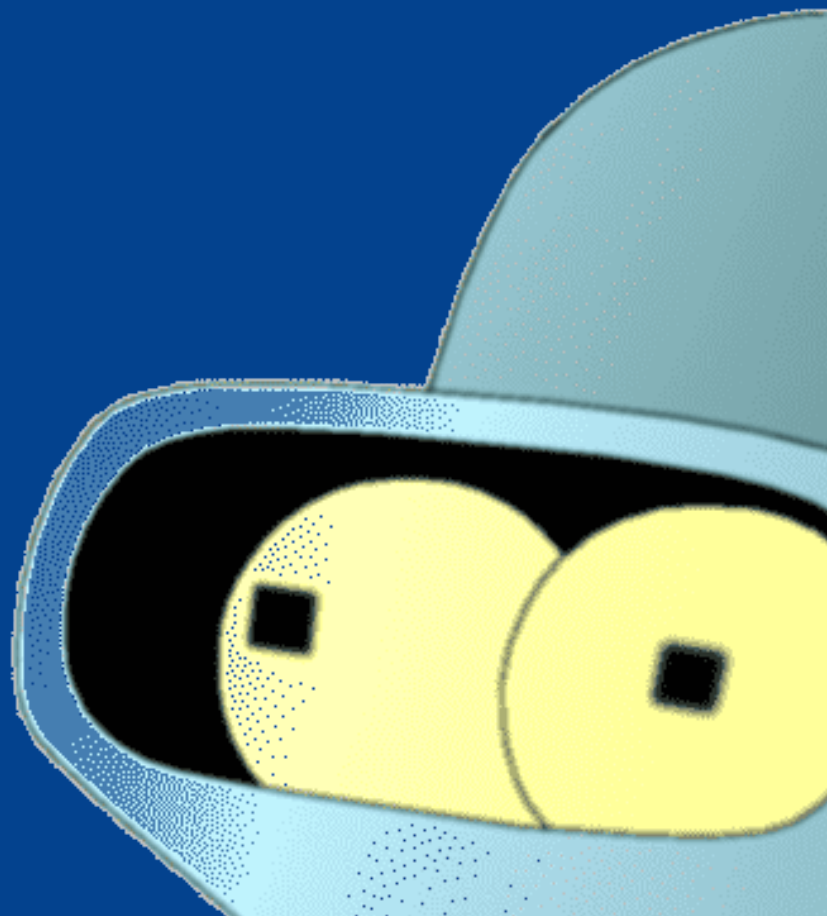
A **large language model** (**LLM**) is a type of computational model designed for natural language processing tasks such as language generation. As language models, LLMs acquire these abilities by learning statistical relationships from vast amounts of text during a self-supervised and semi-supervised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

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.[2] These models acquire predictive power regarding syntax, semantics, and ontologies[3] inherent in human language

Part of a series on
Machine learning and data mining

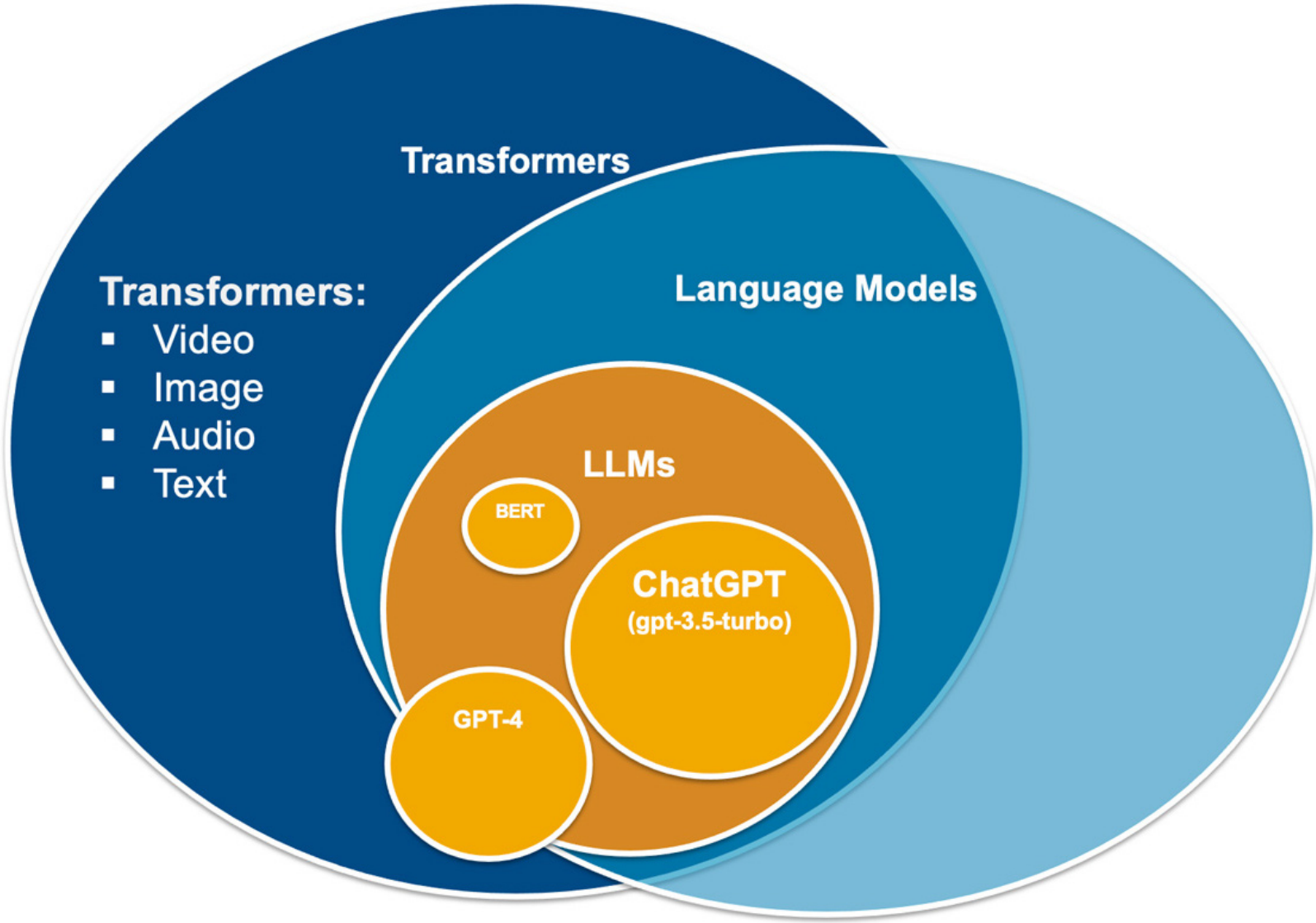
Paradigms	[show]
Problems	[show]
Supervised learning (**classification** • **regression**)	[show]
Clustering	[show]
Dimensionality reduction	[show]
Structured prediction	[show]
Anomaly detection	[show]
Artificial neural network	[hide]
Autoencoder • Deep learning •	

GPT



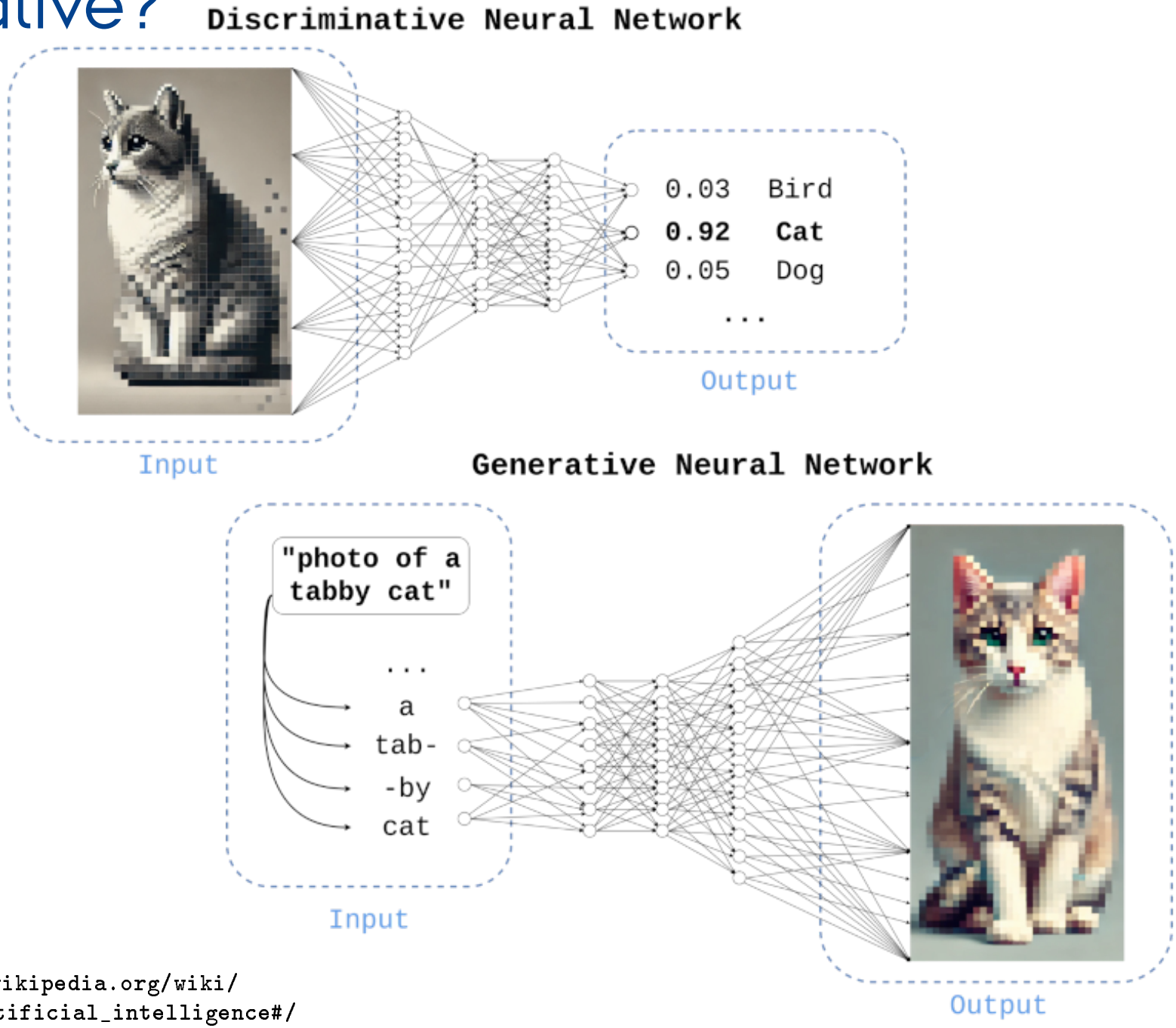
GPT

Generative Pre-trained Transformer



[https://www.mathworks.com/content/dam/mathworks/ mathworks-dot-com/cmsimages/ discovery/images/ chatgpt-discovery-page-llm-transformers-diagram.jpg]

Generative?



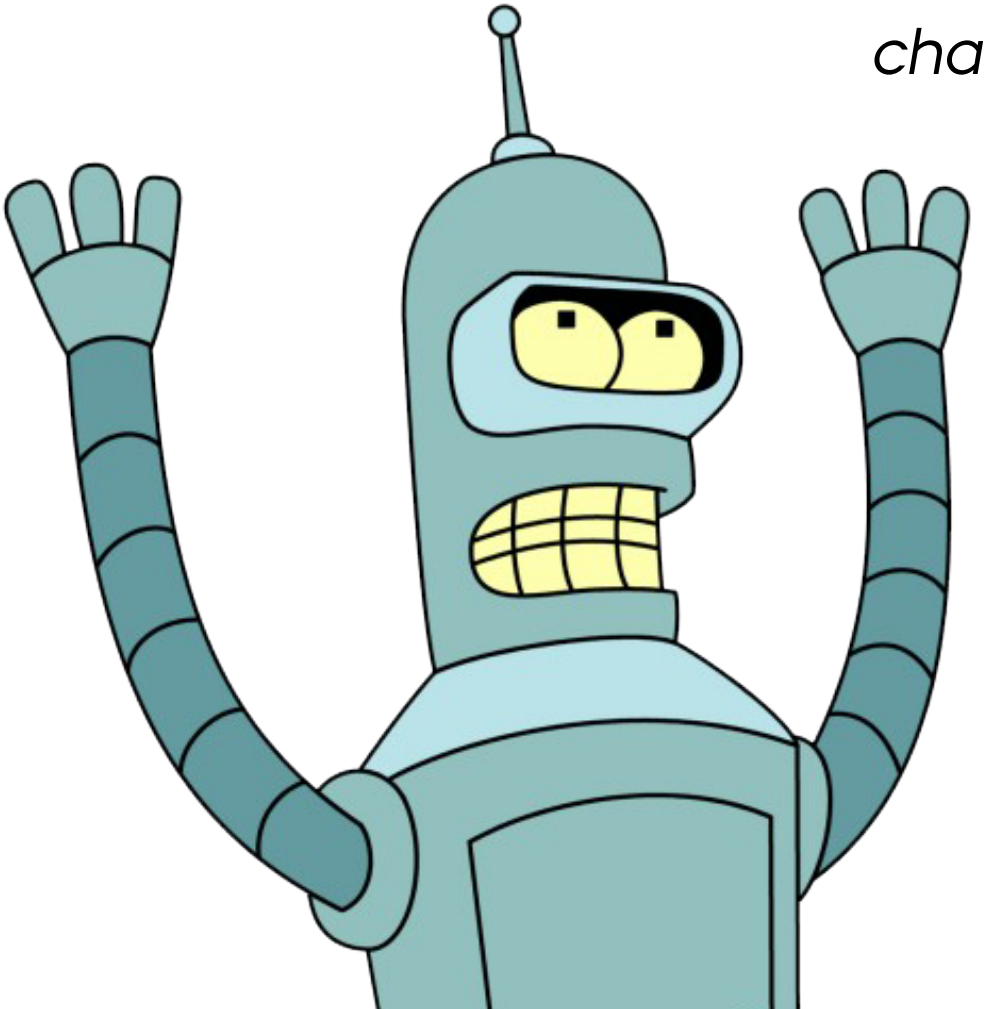
[https://en.wikipedia.org/wiki/Generative_artificial_intelligence#/media/ File:Discriminative_vs_Generative_Neural_Networks.png]

Pre-trained?

..is pre-trained just trained?

..but a Transformer alright!

chat-T, chat-GT, and GAI?



GPT HANDS-ON EXERCISE

UNUSED SLIDES..

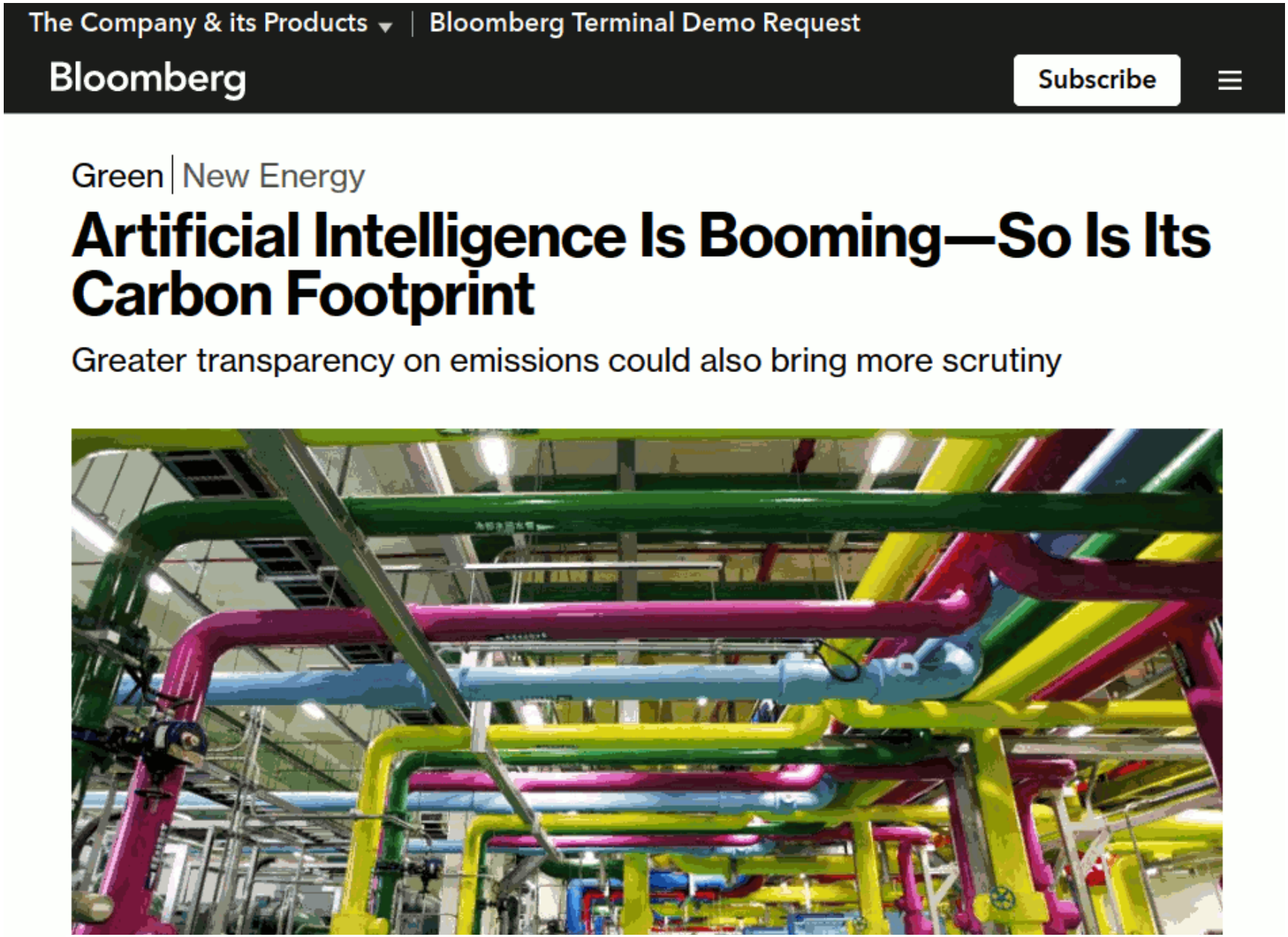
Facts and Reflections

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

Training Cost and Hardware



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