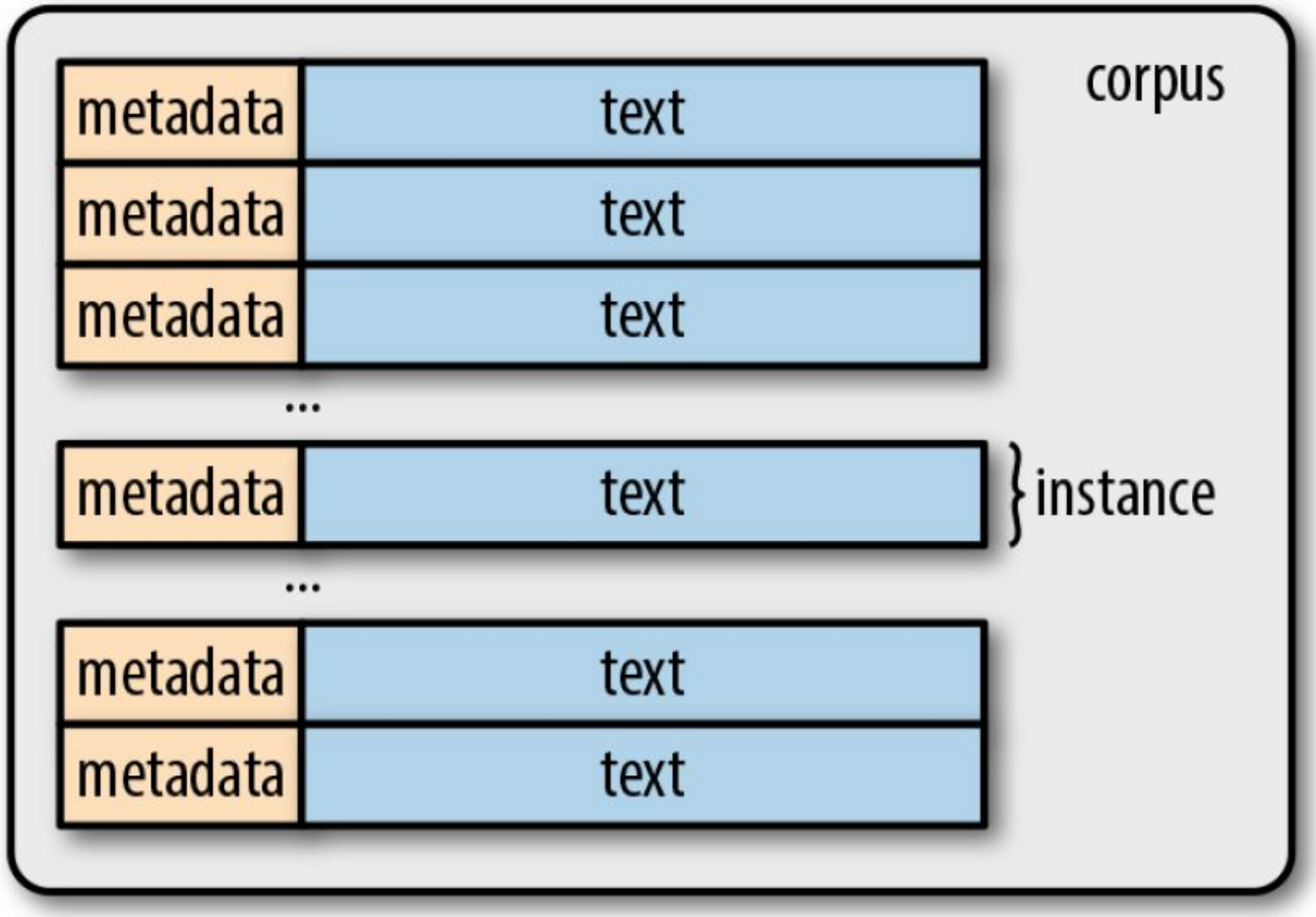
Workshop on
TinyML for
Sustainable Development

Language Models Introduction

Prof. Jesús Alfonso López
jalopez@uao.edu.co
Universidad Autónoma de Occidente

Language Models

Corpus



Corpus = Data Set

Language Models

Token

Raw text is a sequence of characters (bytes), but most of the time it is useful to group the characters into contiguous units called tokens. Tokens correspond to words, parts of words, and numerical sequences separated by white space or punctuation marks.

Input[0]

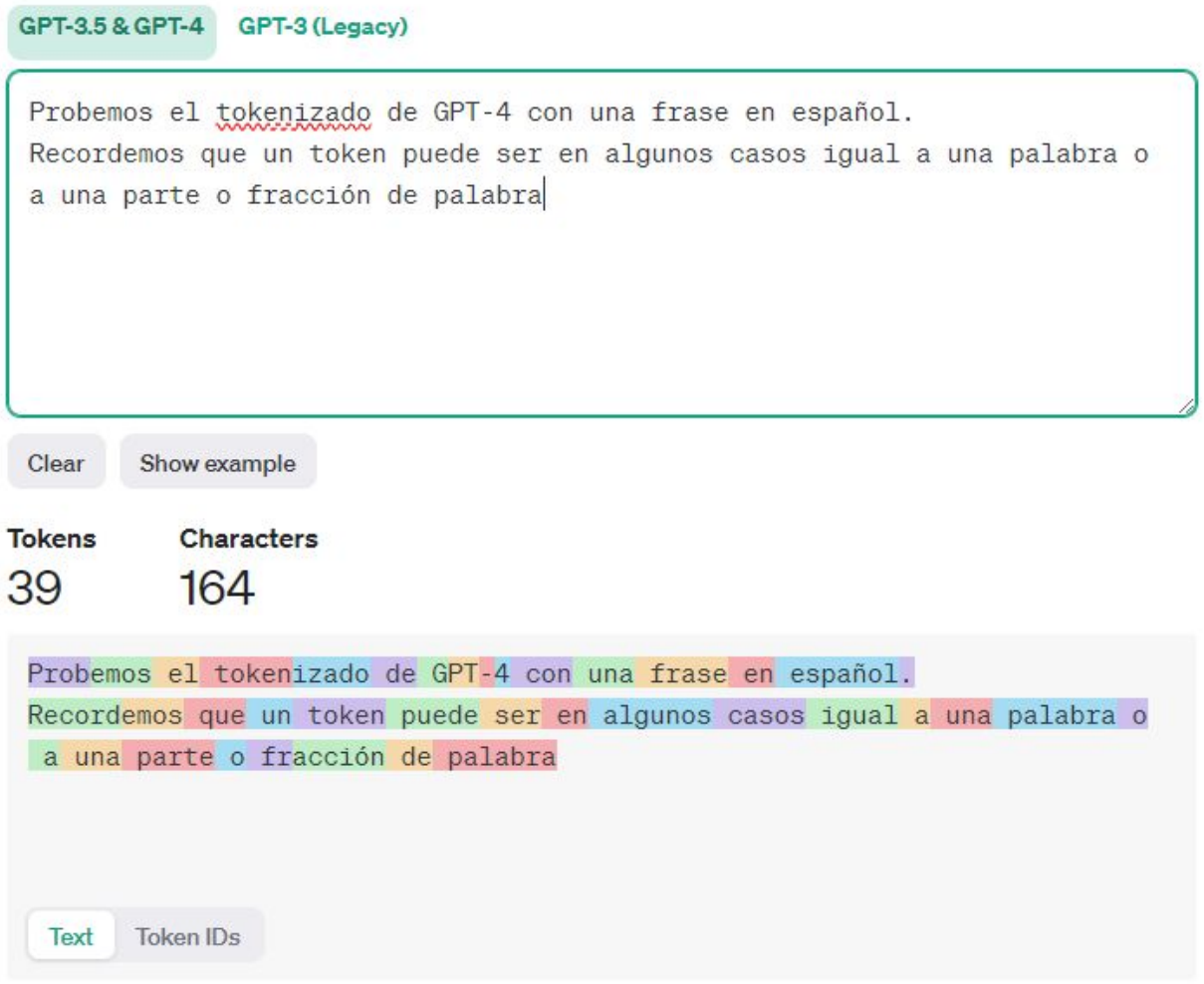
```
import spacy
nlp = spacy.load('en')
text = "Mary, don’t slap the green witch"
print([str(token) for token >in nlp(text.lower())])
```

Output[0]

```
['mary', ',', 'do', "n't", 'slap', 'the', 'green', 'witch', '.']
```

Language Models

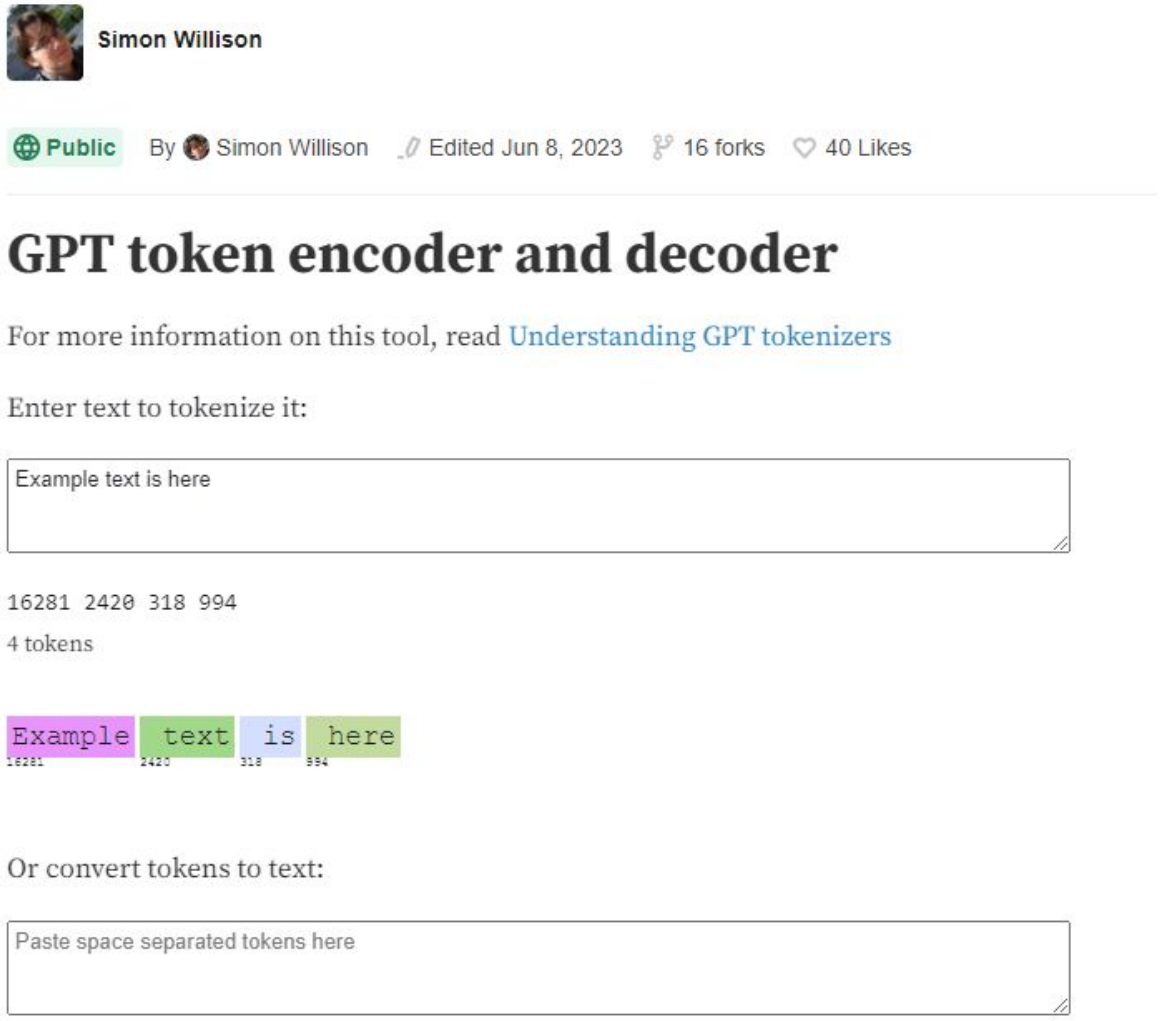
Token



https://platform.openai.com/tokenizer

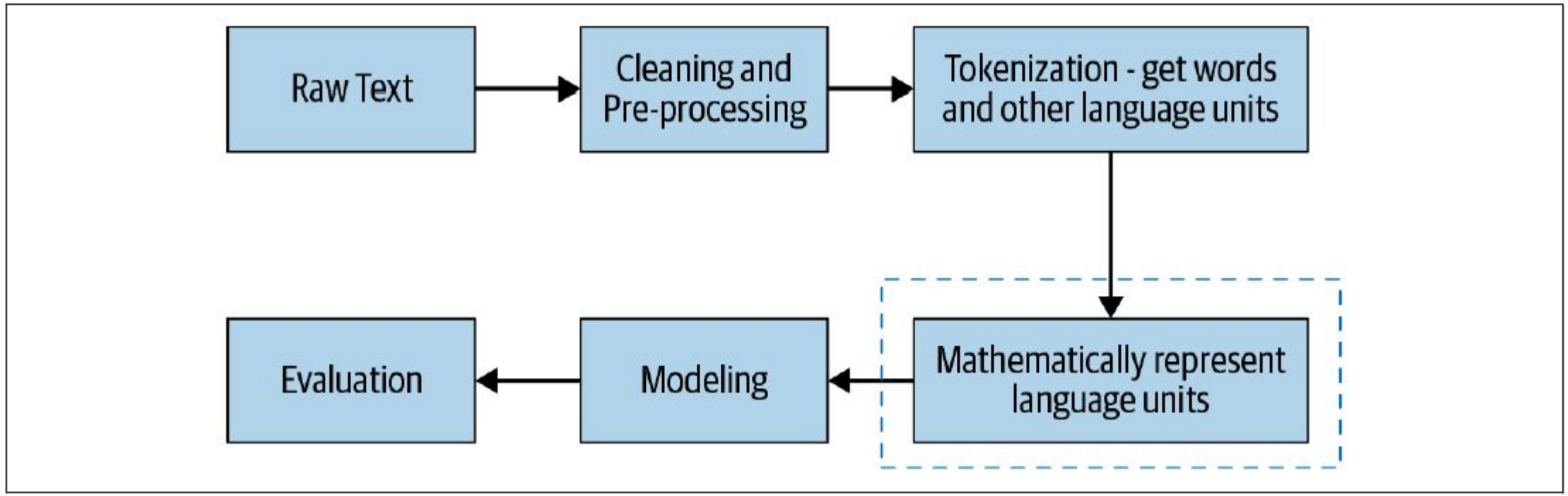
Language Models

Token

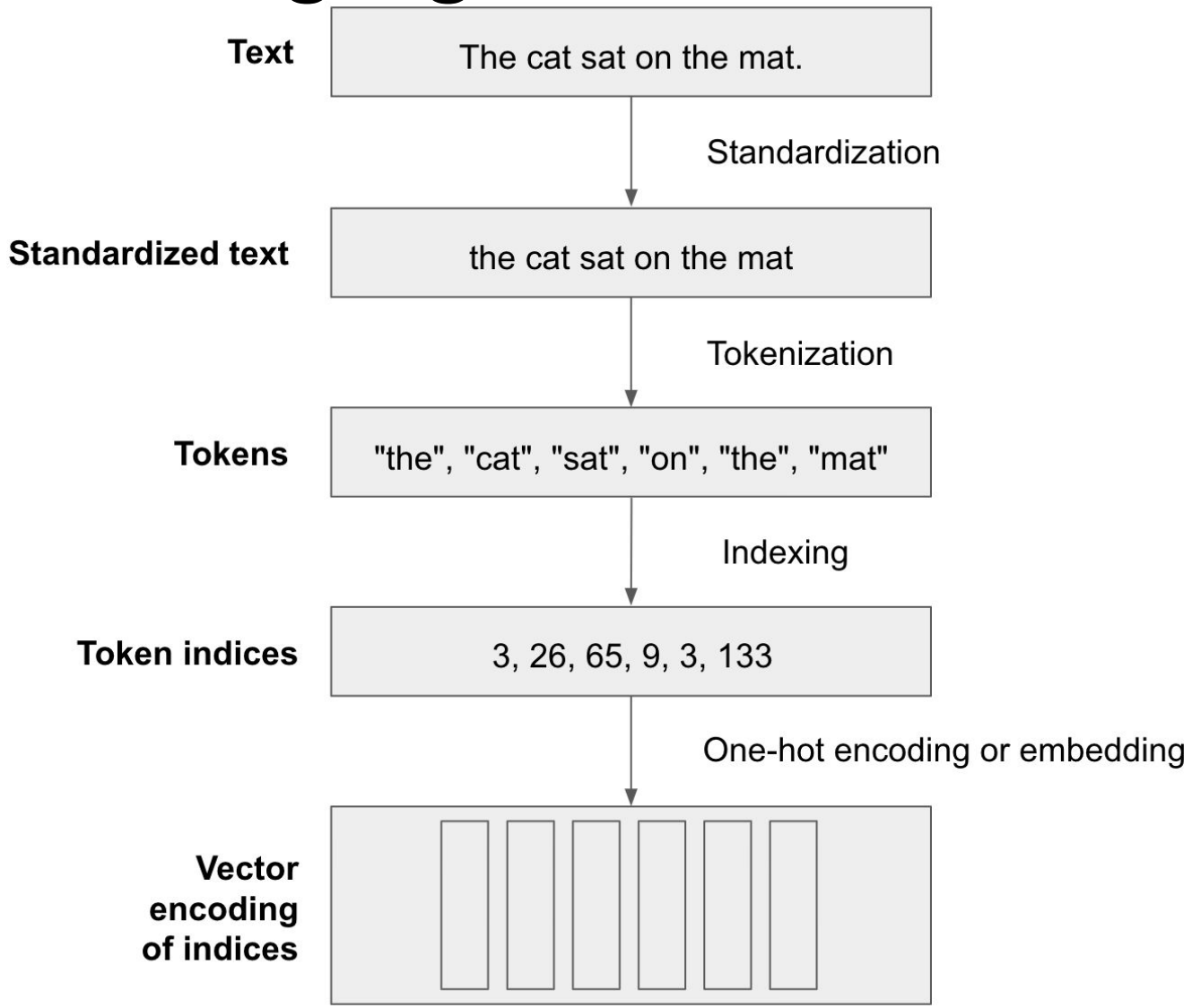


https://observablehq.com/@simonw/gpt-tokenizer

Language Models



Language Models



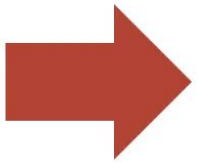
https://livebook.manning.com/book/deep-learning-with-python-second-edition/chapter-11/v-6/28

Language Models

Word embedding = From tokens to vectors

One hot encoding

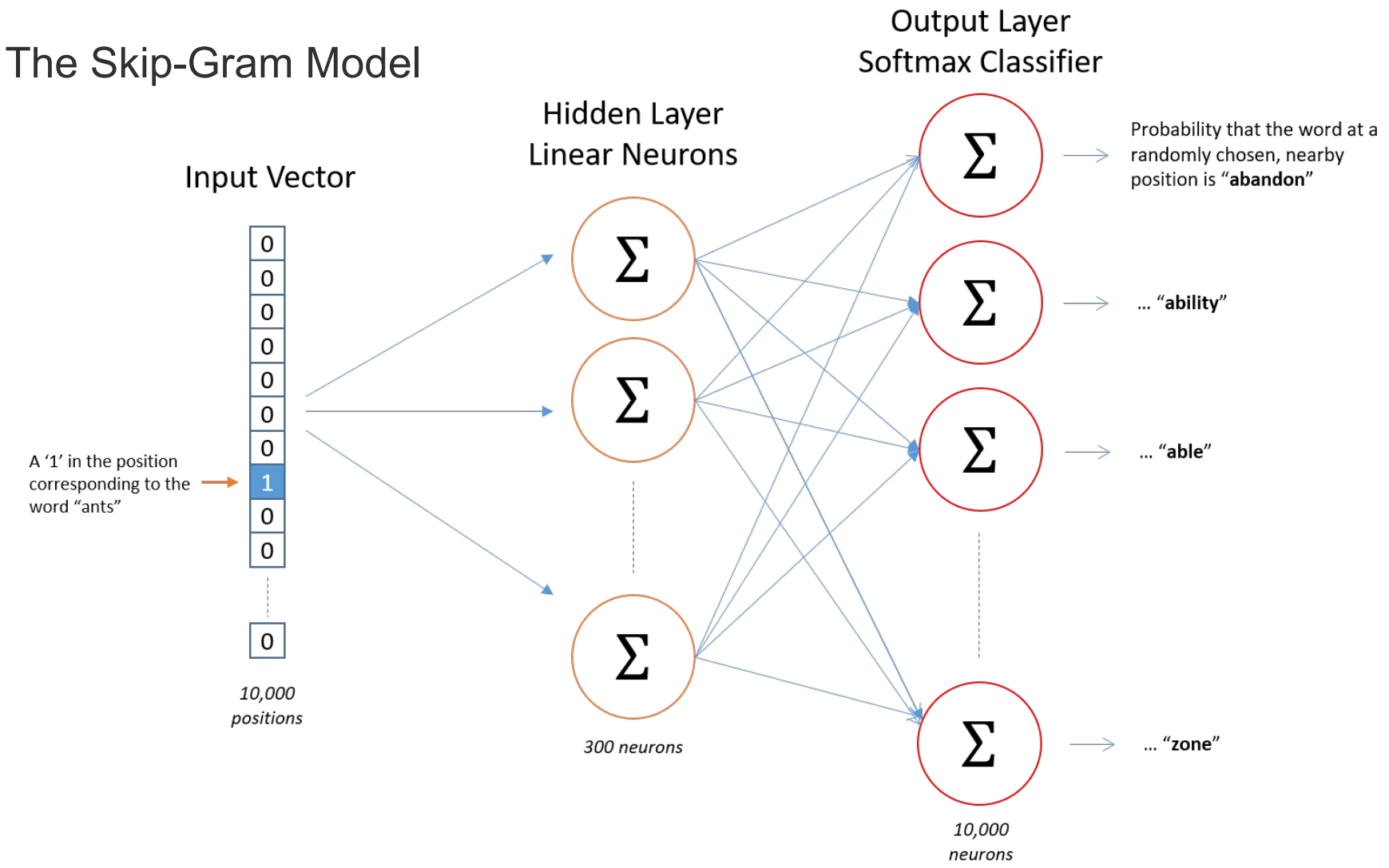
Vocabulary:
Man, woman, boy, girl, prince, princess, queen, king, monarch



	1	2	3	4	5	6	7	8	9
man	1	0	0	0	0	0	0	0	0
woman	0	1	0	0	0	0	0	0	0
boy	0	0	1	0	0	0	0	0	0
girl	0	0	0	1	0	0	0	0	0
prince	0	0	0	0	1	0	0	0	0
princess	0	0	0	0	0	1	0	0	0
queen	0	0	0	0	0	0	1	0	0
king	0	0	0	0	0	0	0	1	0
monarch	0	0	0	0	0	0	0	0	1

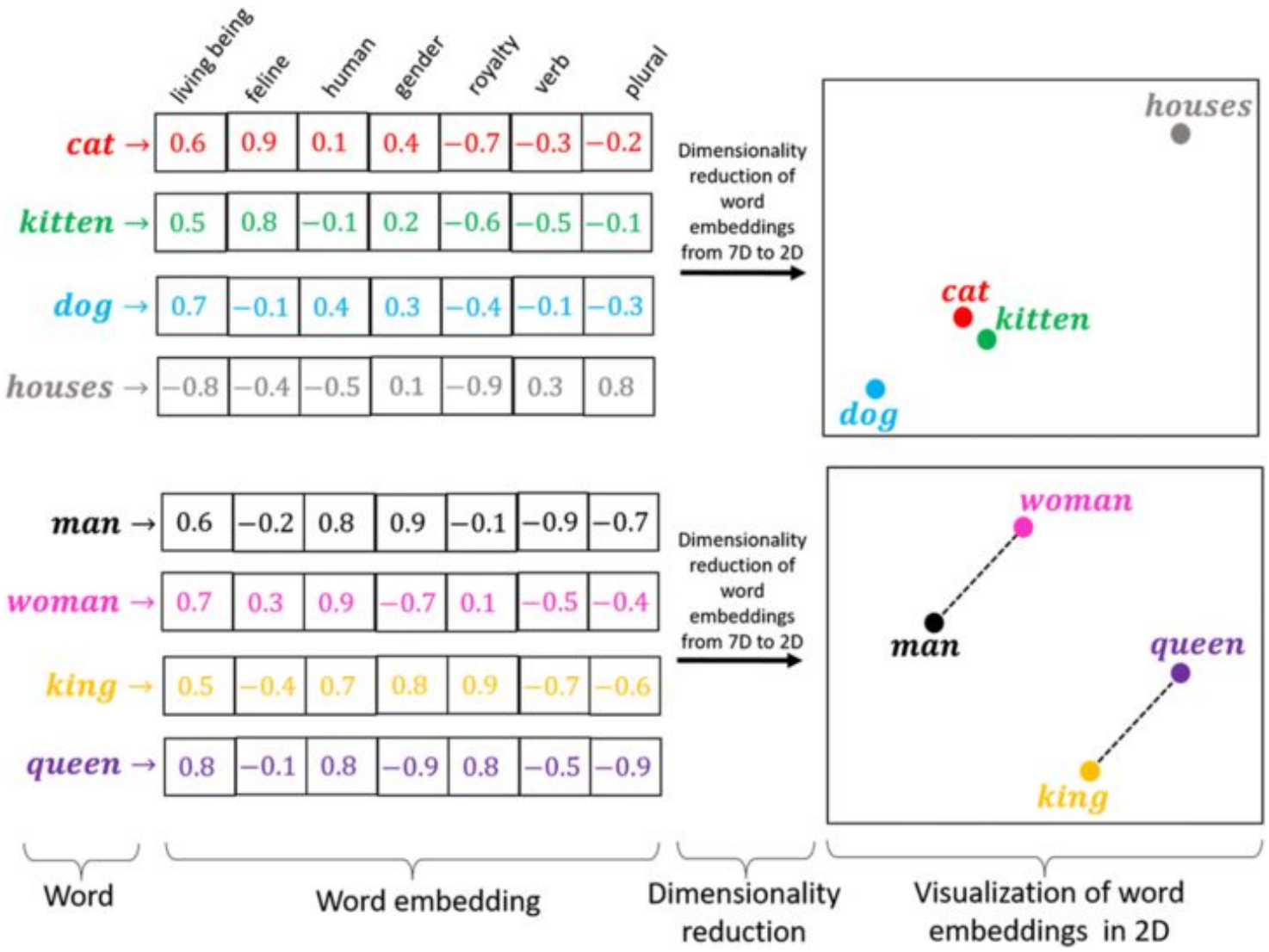
Each word gets a 1x9 vector representation

Language Models

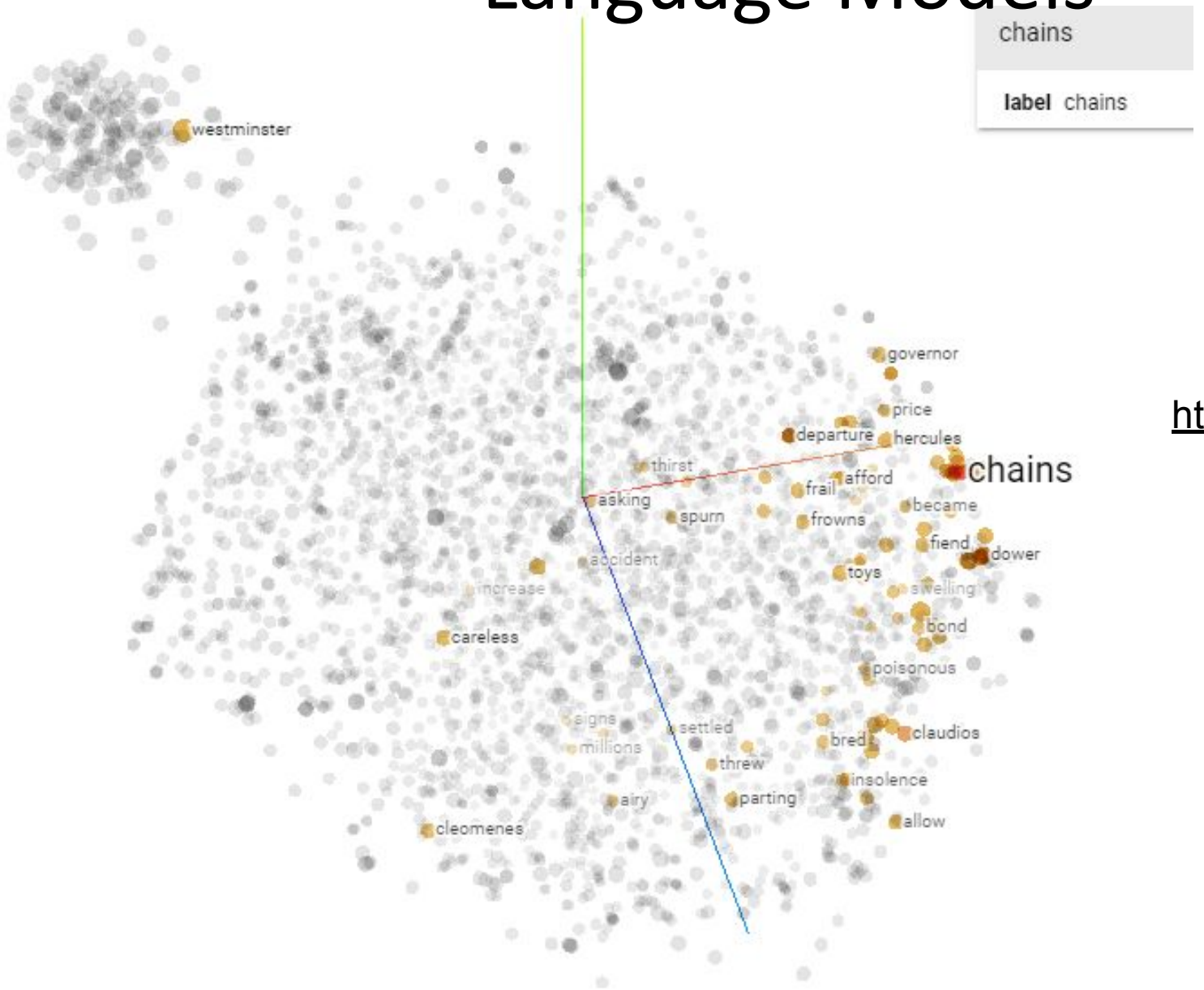


http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

Language Models

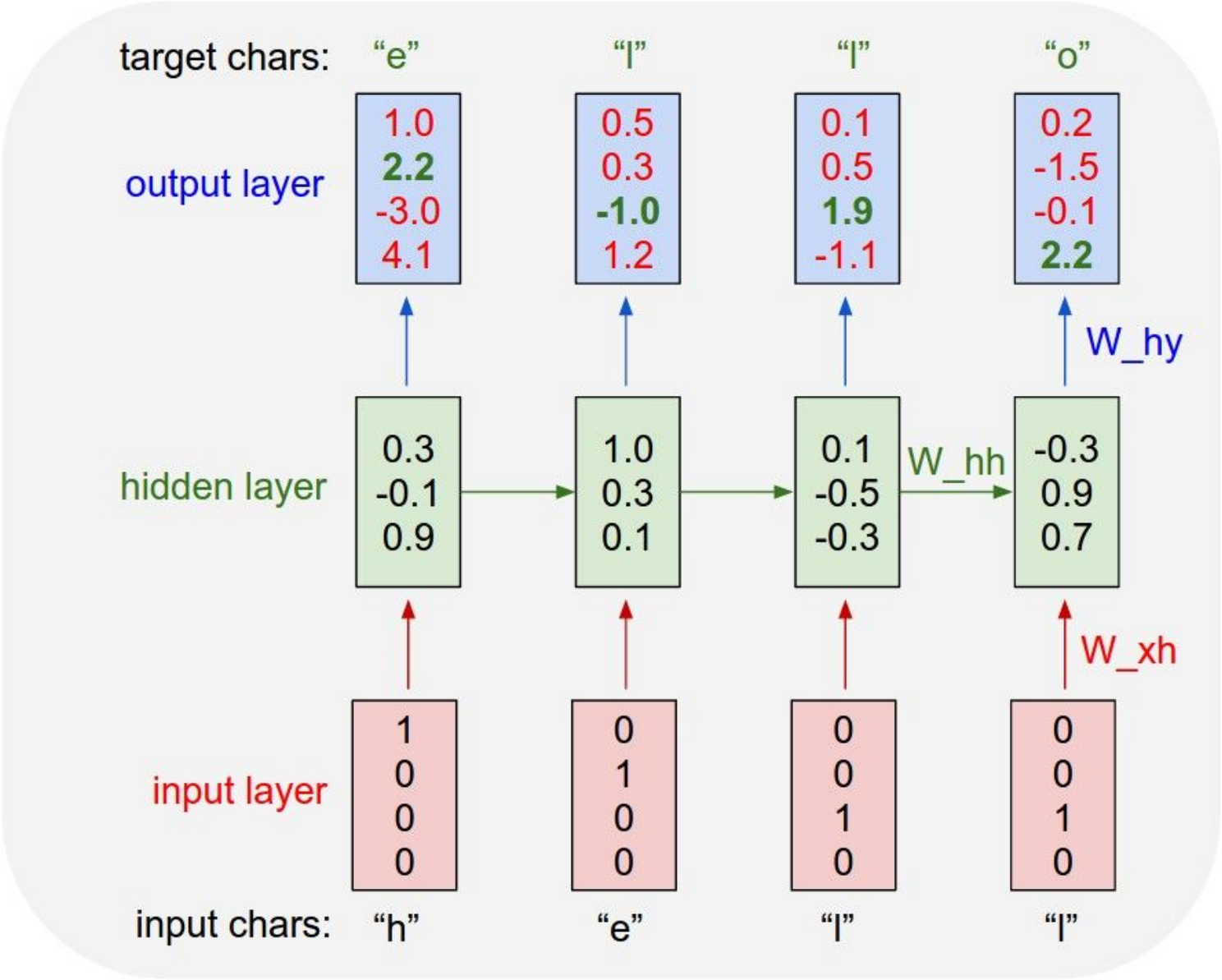


Language Models



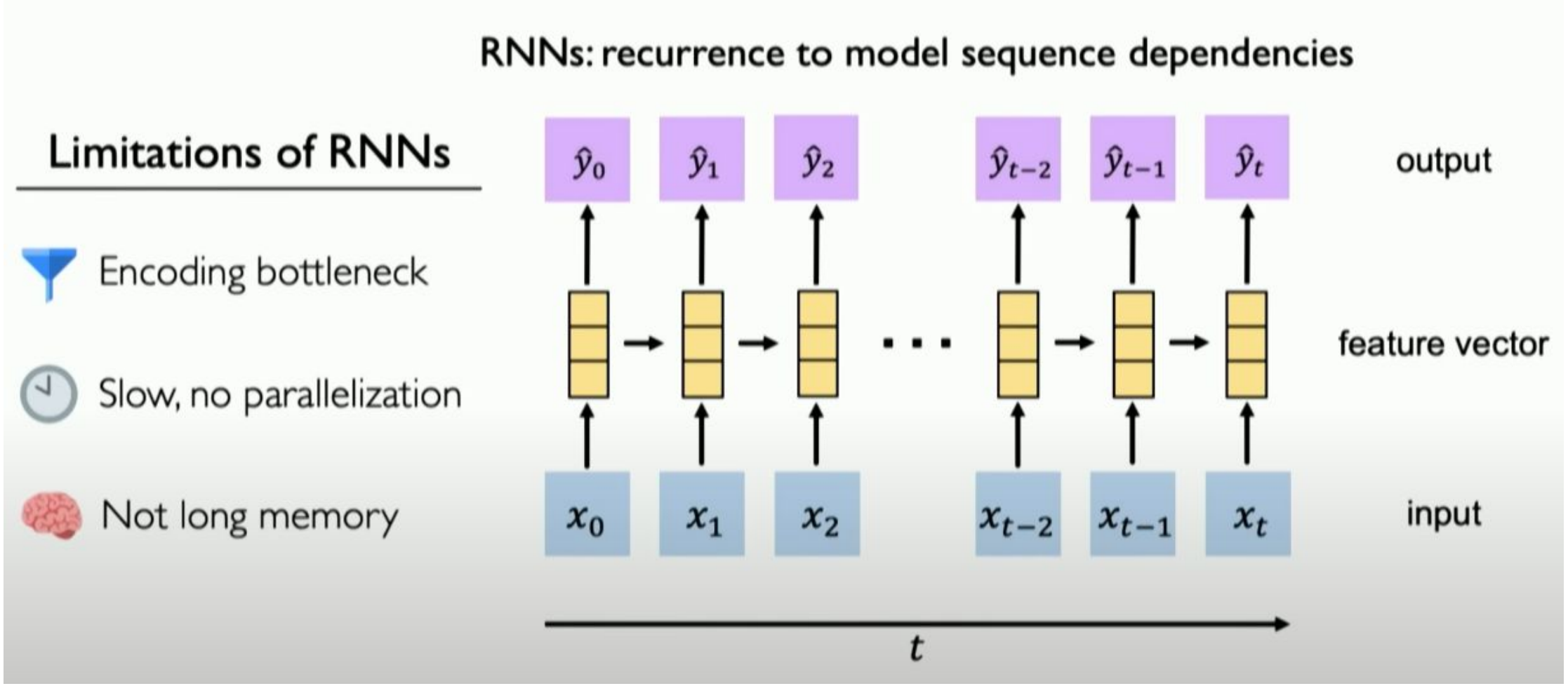
https://projector.tensorflow.org/

Language Models



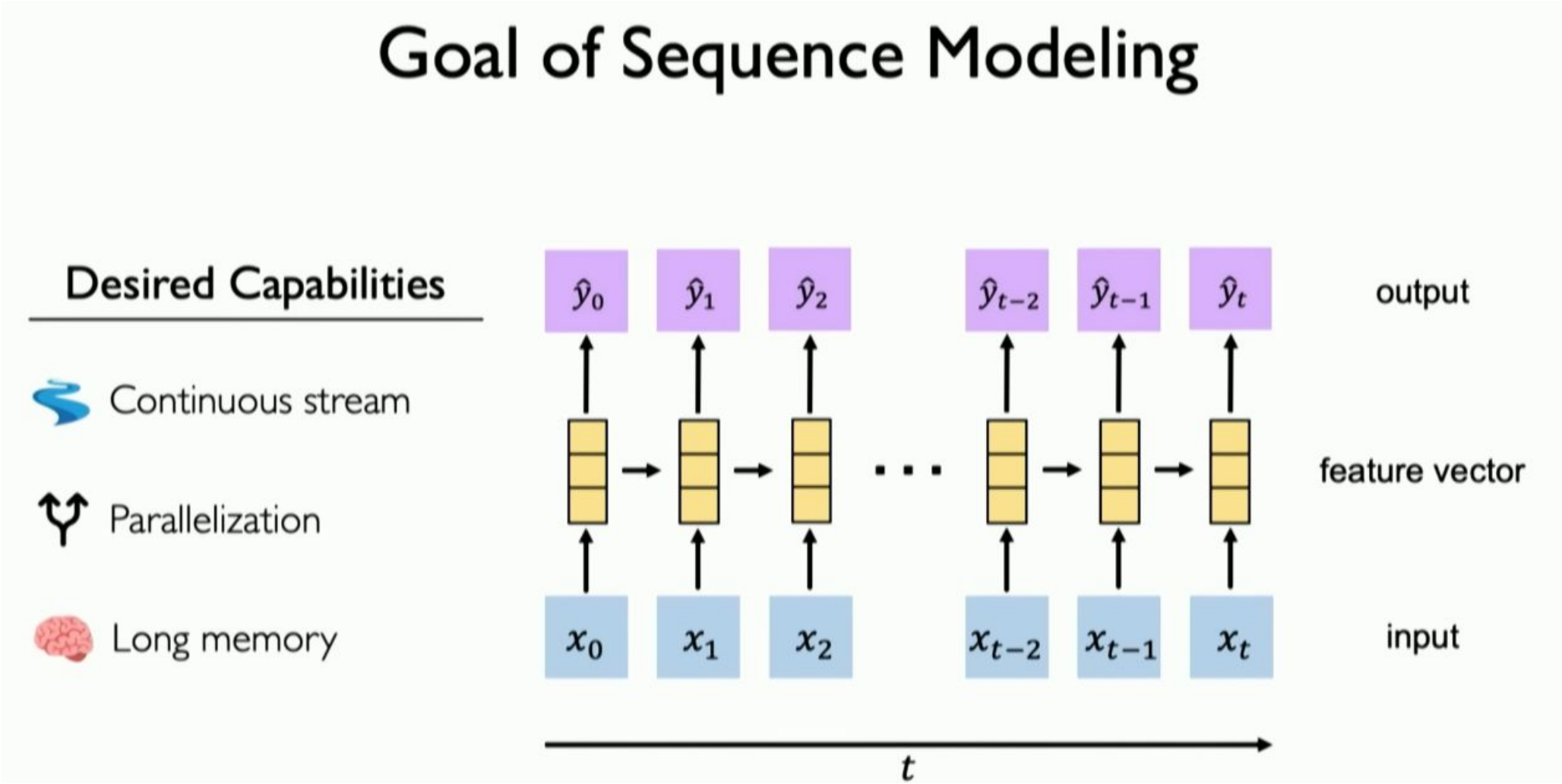
https://karpathy.github.io/2015/05/21/rnn-effectiveness/

Language Models



https://www.youtube.com/watch?v=ySEx_Bqxvvo

Language Models



https://www.youtube.com/watch?v=ySEx_Bqxvvo

Language Models

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

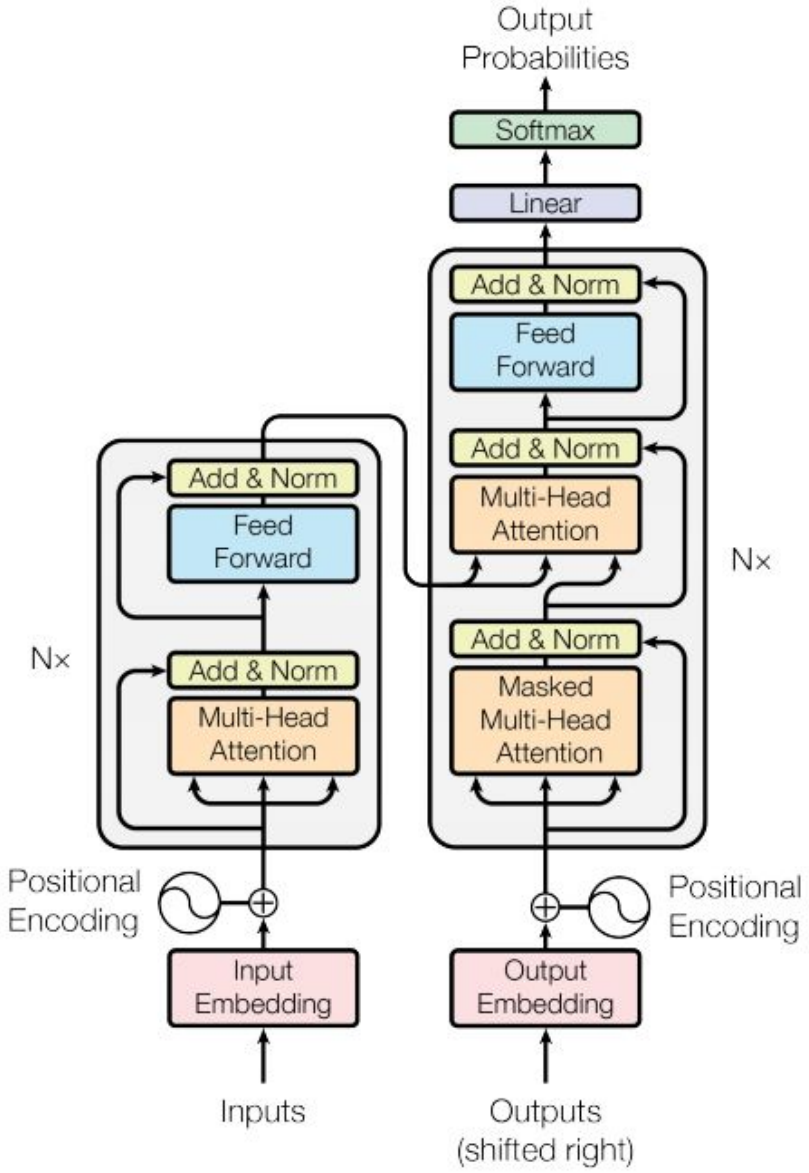
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

https://arxiv.org/pdf/1706.03762.pdf

Transformer

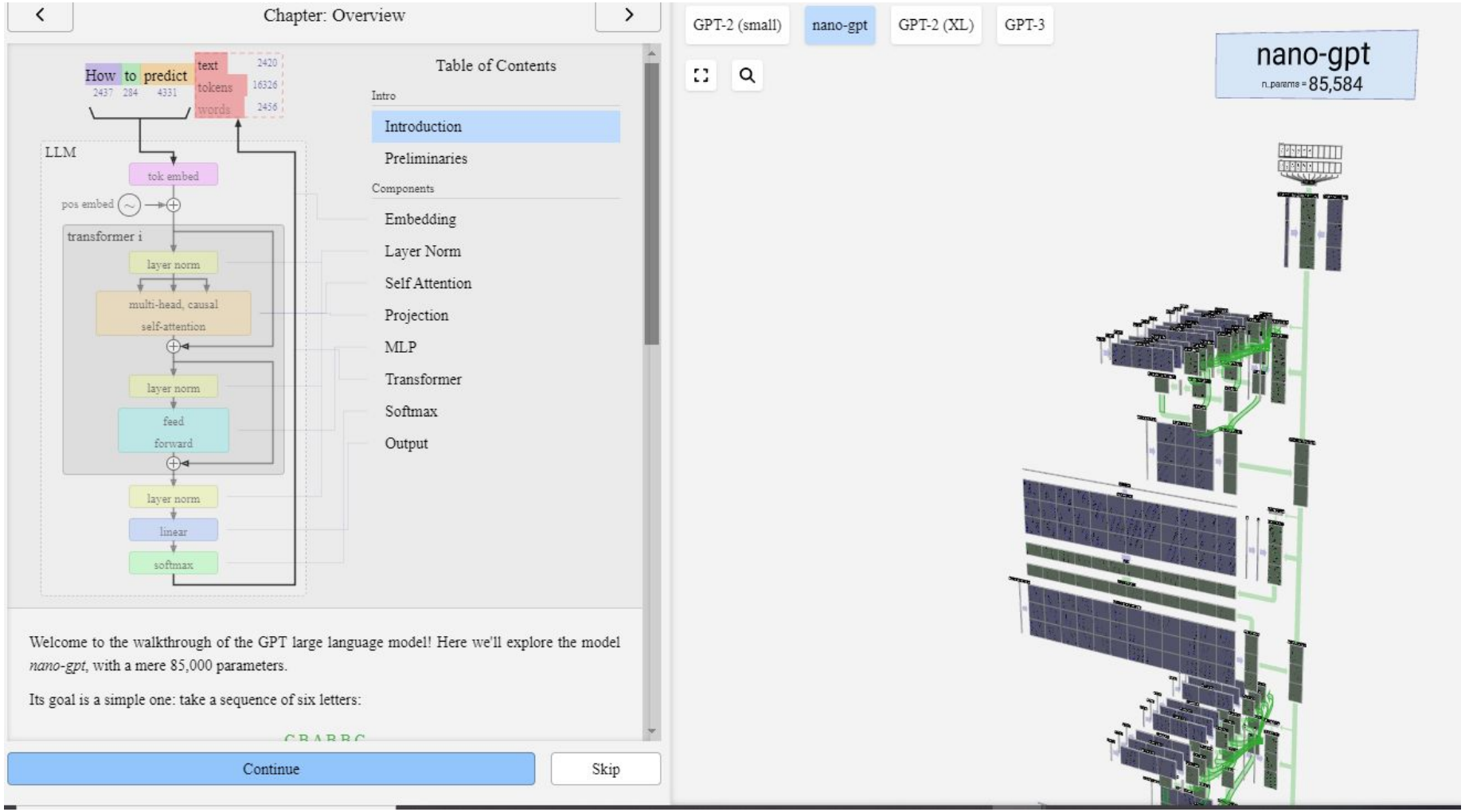


Language Models

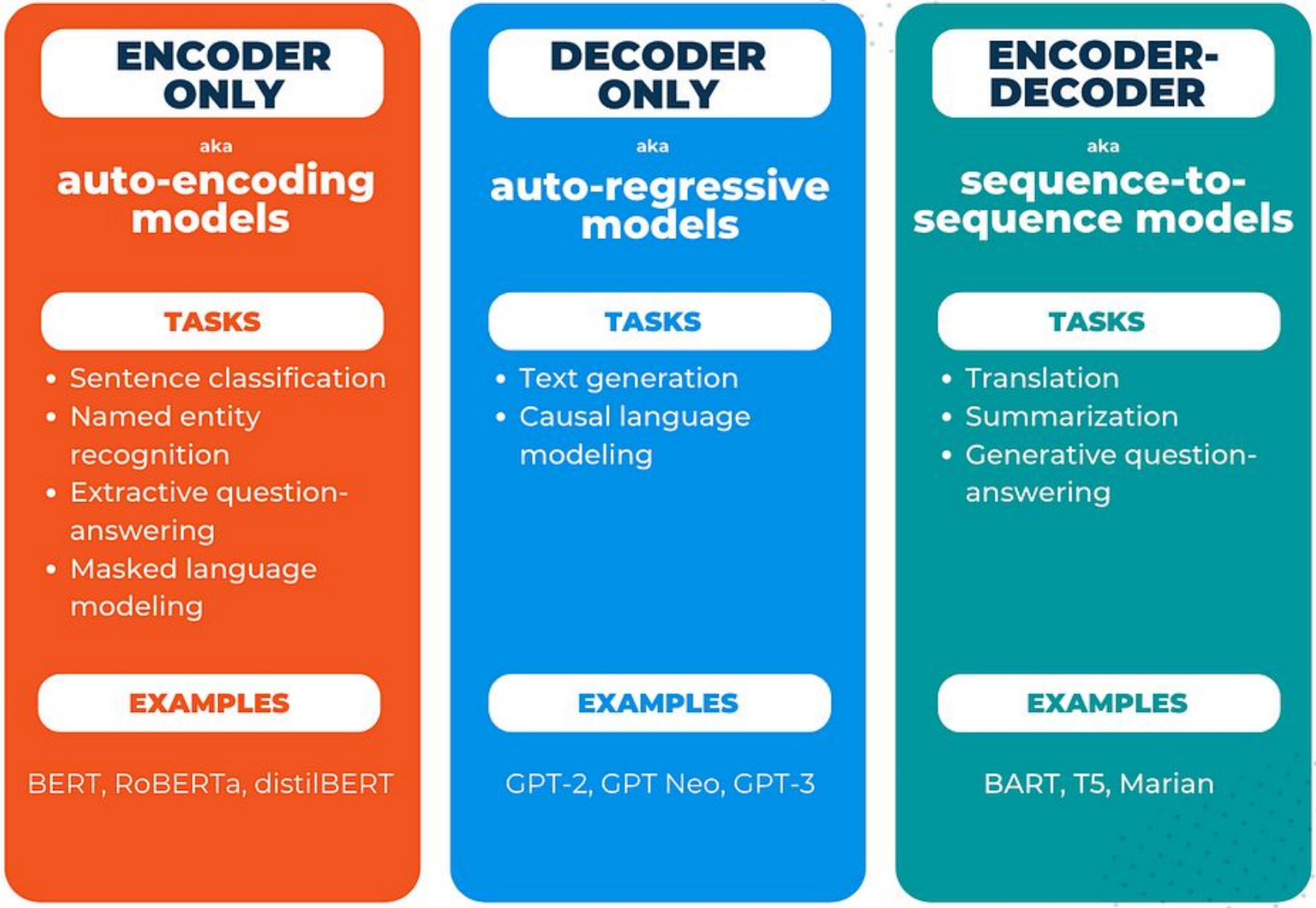
https://bbycroft.net/llm

http://jalammar.github.io/illustrated-transformer/

https://www.youtube.com/watch?v=wjZofJX0v4M



Language Models



https://mlops.community/explainable-ai-visualizing-attention-in-transformers/

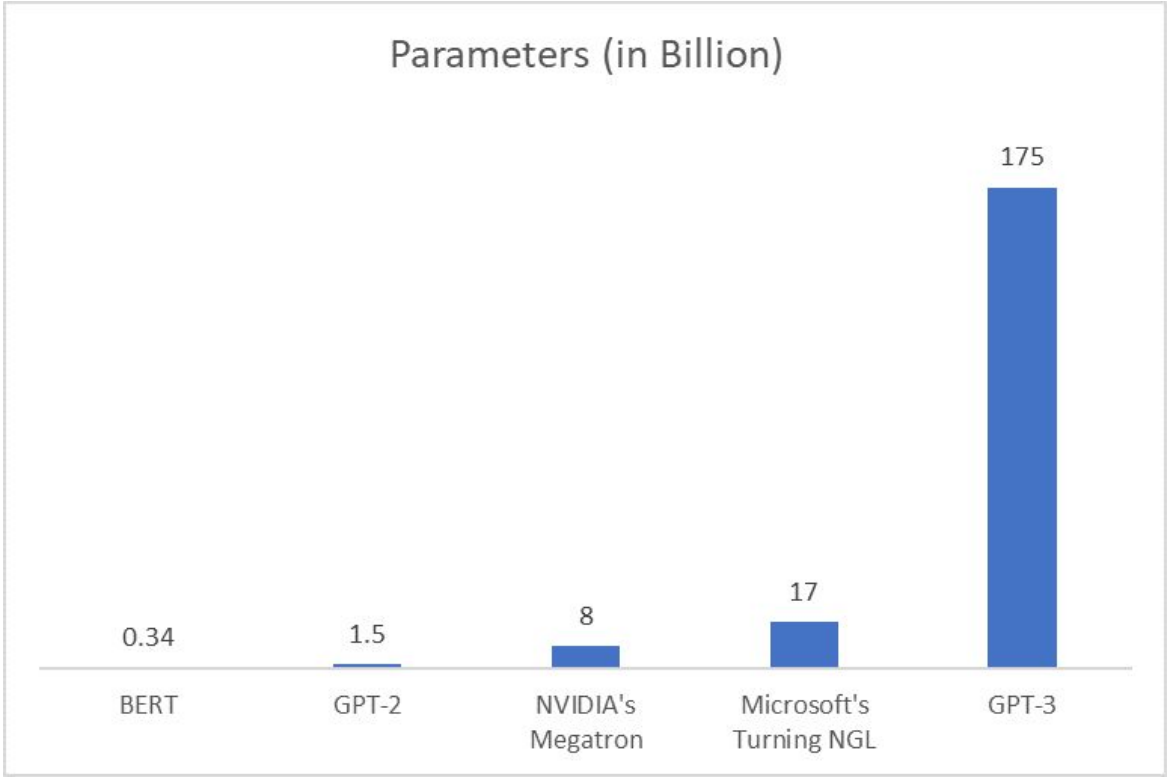
Language Models

Grandes Modelos de lenguaje (Large Language Model)

Year	Model	# of Parameters	Dataset Size
2019	BERT [39]	3.4E+08	16GB
2019	DistilBERT [113]	6.60E+07	16GB
2019	ALBERT [70]	2.23E+08	16GB
2019	XLNet (Large) [150]	3.40E+08	126GB
2020	ERNIE-GEN (Large) [145]	3.40E+08	16GB
2019	RoBERTa (Large) [74]	3.55E+08	161GB
2019	MegatronLM [122]	8.30E+09	174GB
2020	T5-11B [107]	1.10E+10	745GB
2020	T-NLG [112]	1.70E+10	174GB
2020	GPT-3 [25]	1.75E+11	570GB
2020	GShard [73]	6.00E+11	–
2021	Switch-C [43]	1.57E+12	745GB

2021 WuDao 2.0 1.75E+12

2021 The Megatron-Turing Natural Language Generation model (MT-NLG) 5.30 E+11

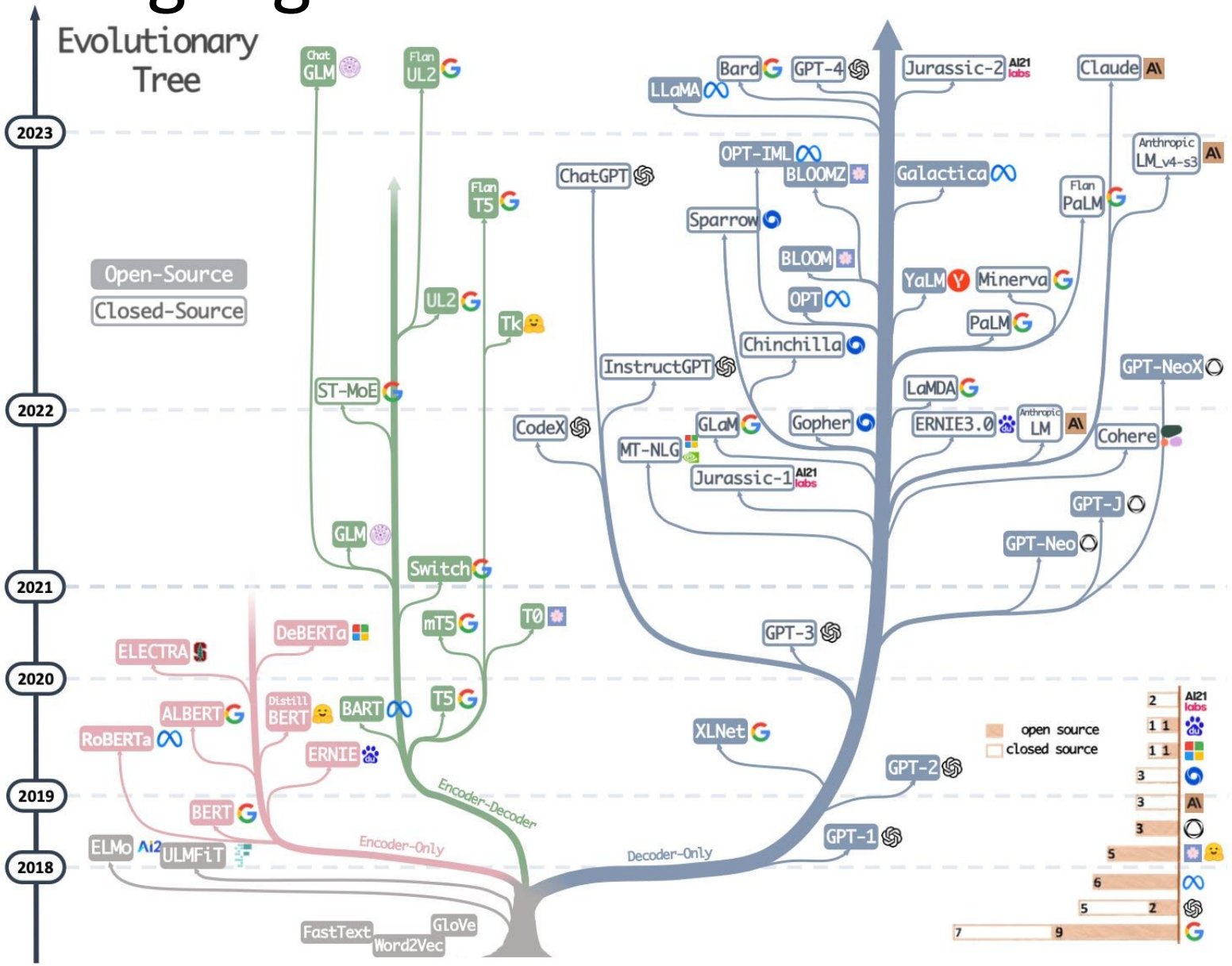


http://faculty.washington.edu/ebender/papers/Stochastic_Parrots.pdf

https://www.merkleinc.com/in/blog/ai-search-what-openais-gpt-3-means-google-and-seo-0

Language Models

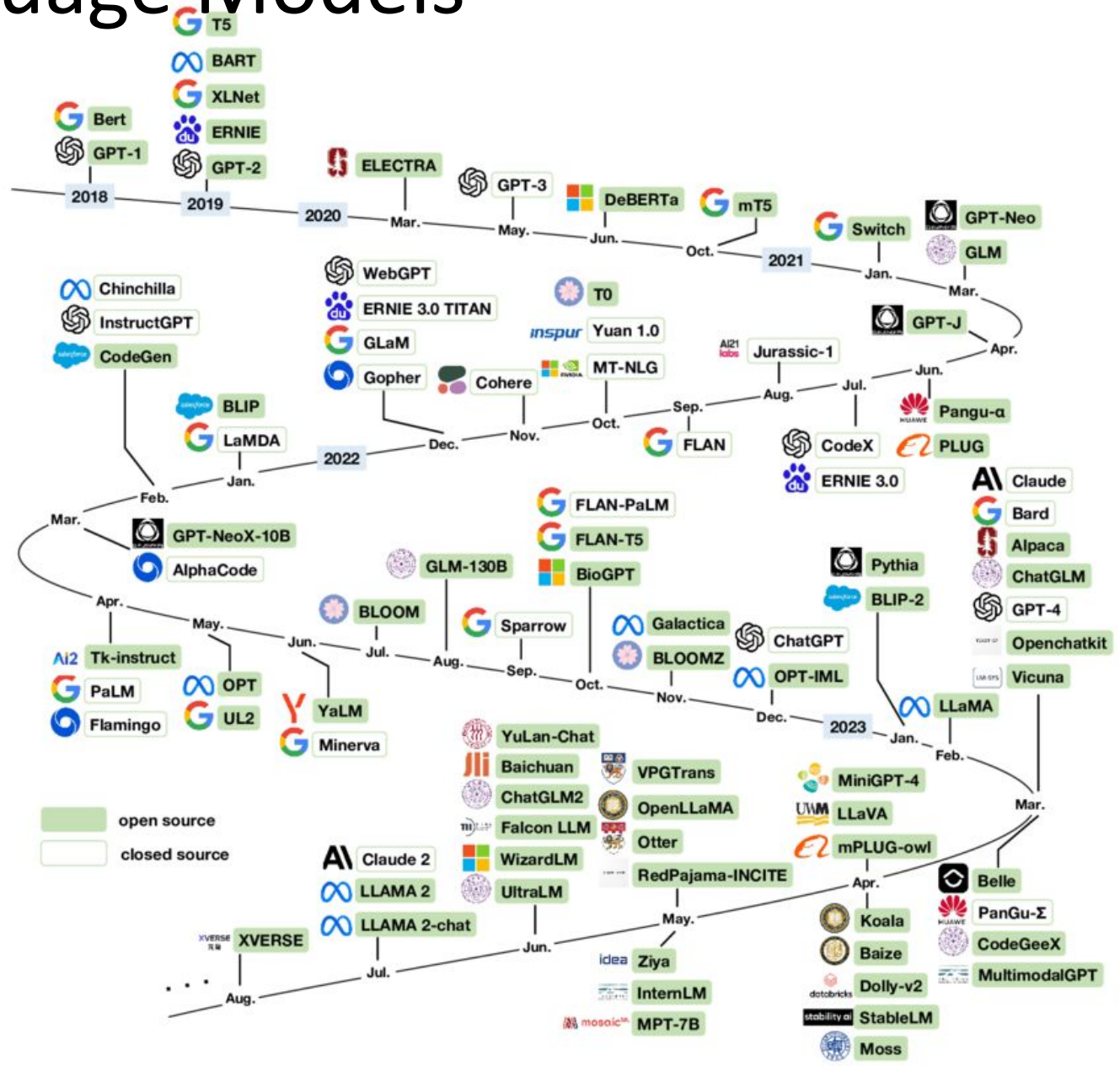
LLM
Evolution



https://magazine.sebastianraschka.com/p/understanding-large-language-models

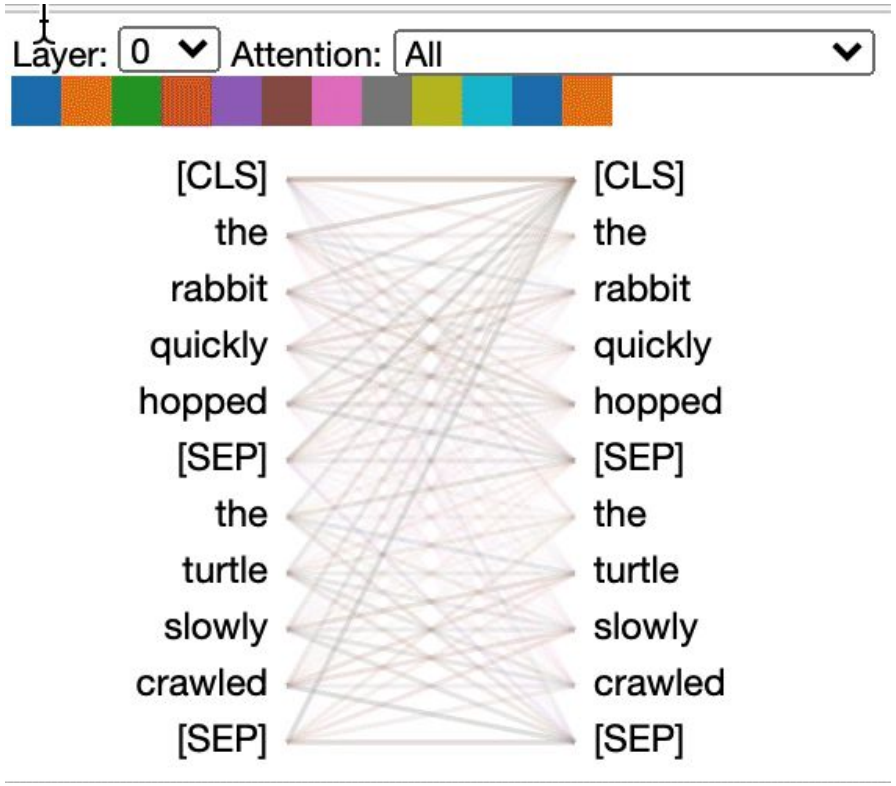
Language Models

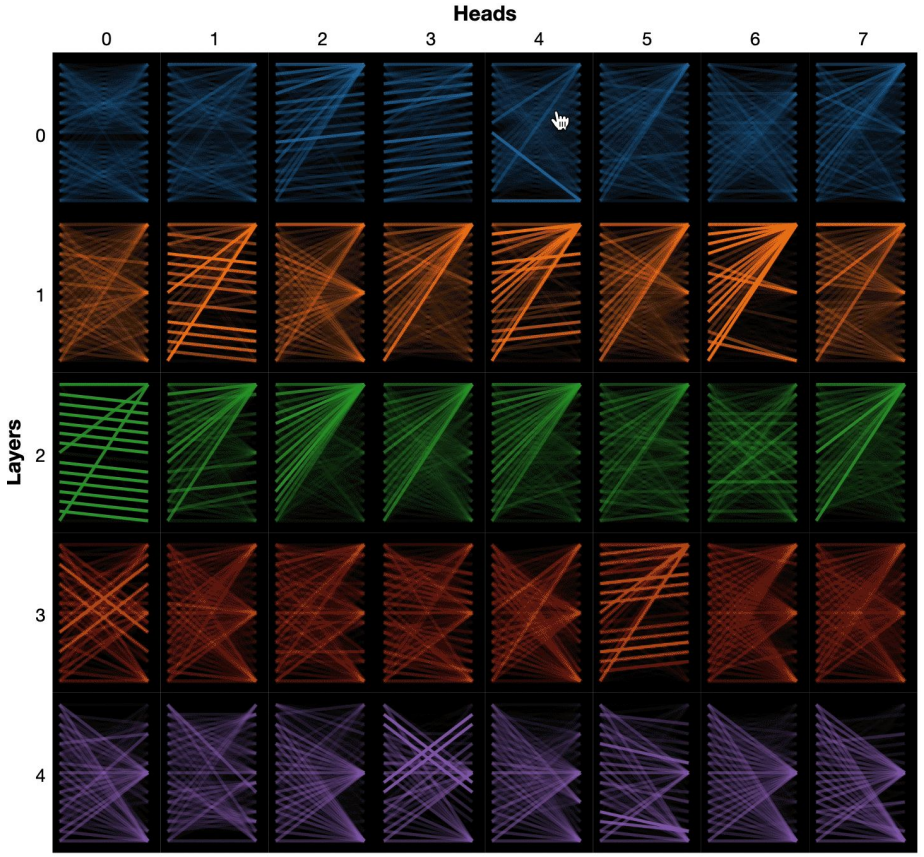
LLM
Evolution



https://www.researchgate.net/figure/A-chronological-overview-of-large-language-models-LLMs-multimodal-and-scientific_fig2_373451304

Language Models





https://github.com/jessevig/bertviz

https://colab.research.google.com/drive/1hXIQ77A4TYS4y3UthWF-Ci7V7vVUoxmQ

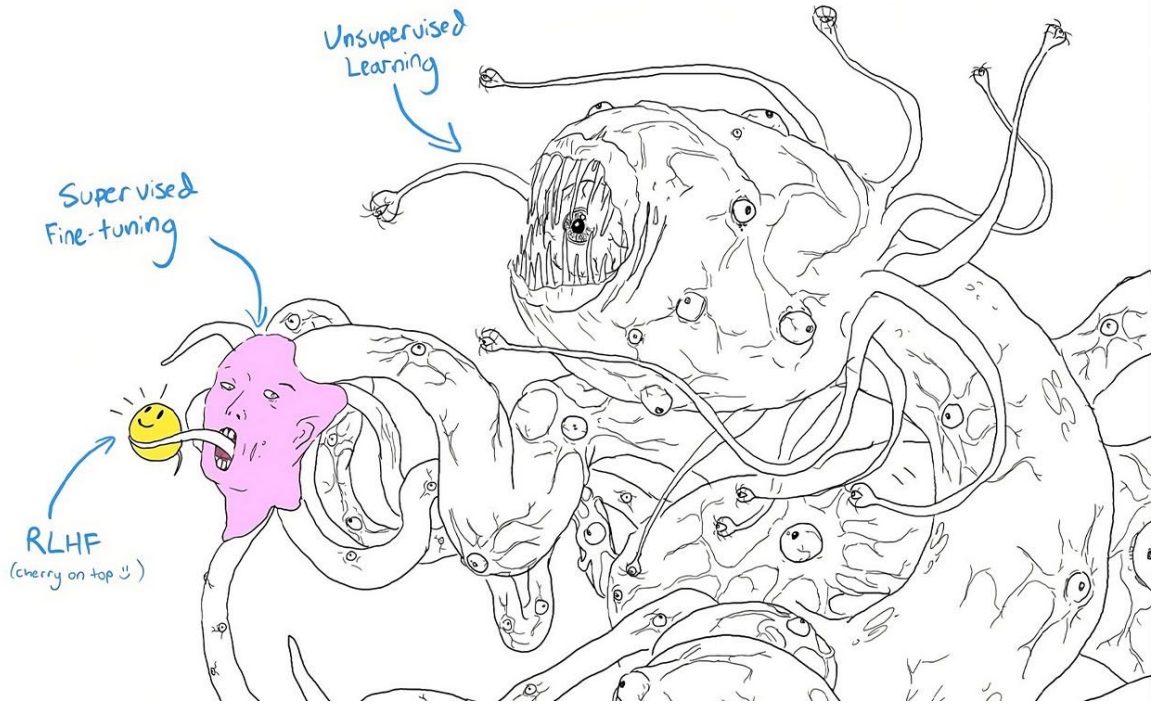
Language Models

https://chatgpt.com/

Language Models

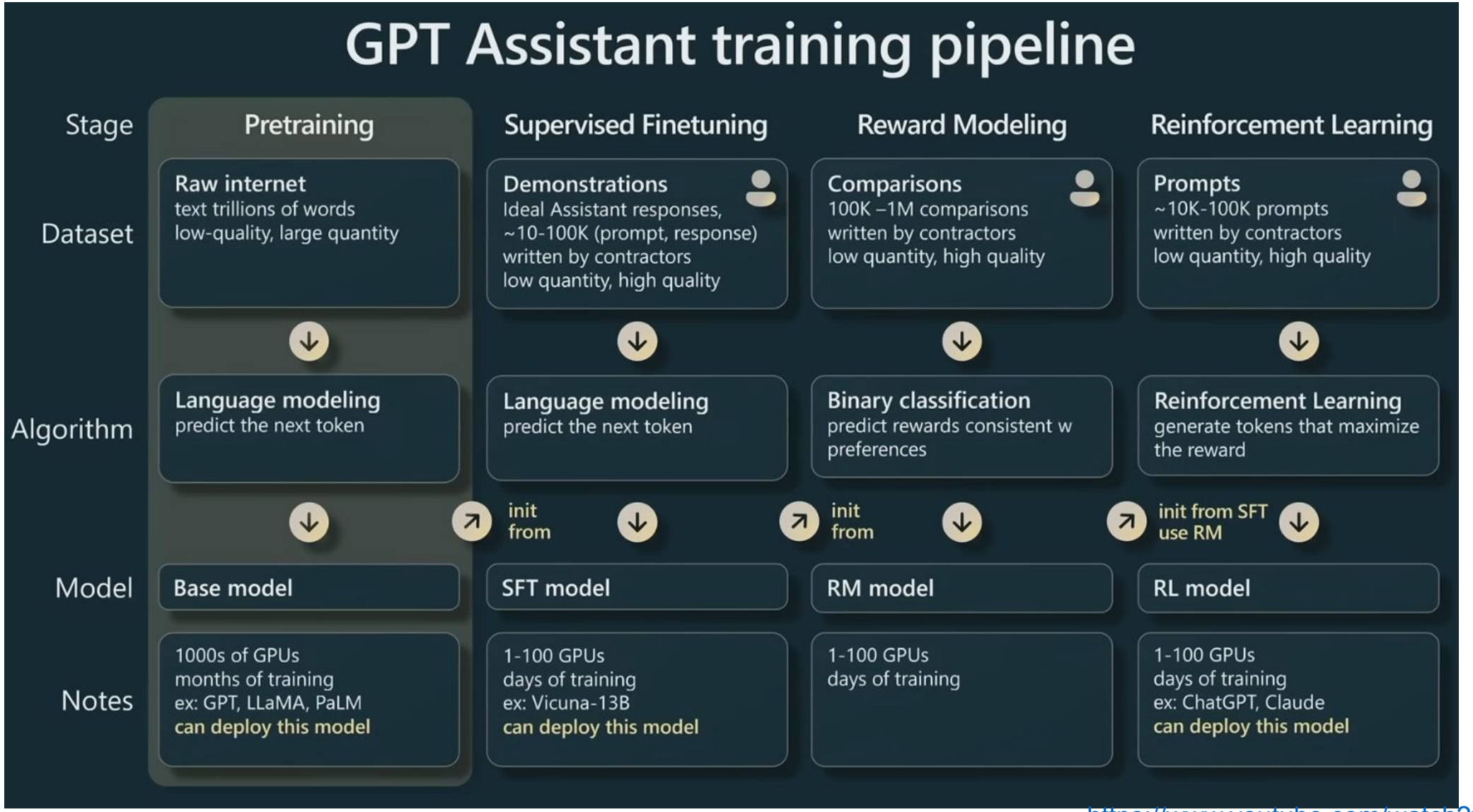
Reinforcement learning from Human Feedback

“Models trained by RLHF can provide answers that align with human values, generate more detailed answers, and reject questions that are inappropriate or outside the model's knowledge space.” Therefore, they can be used to reduce response bias. of the LLM



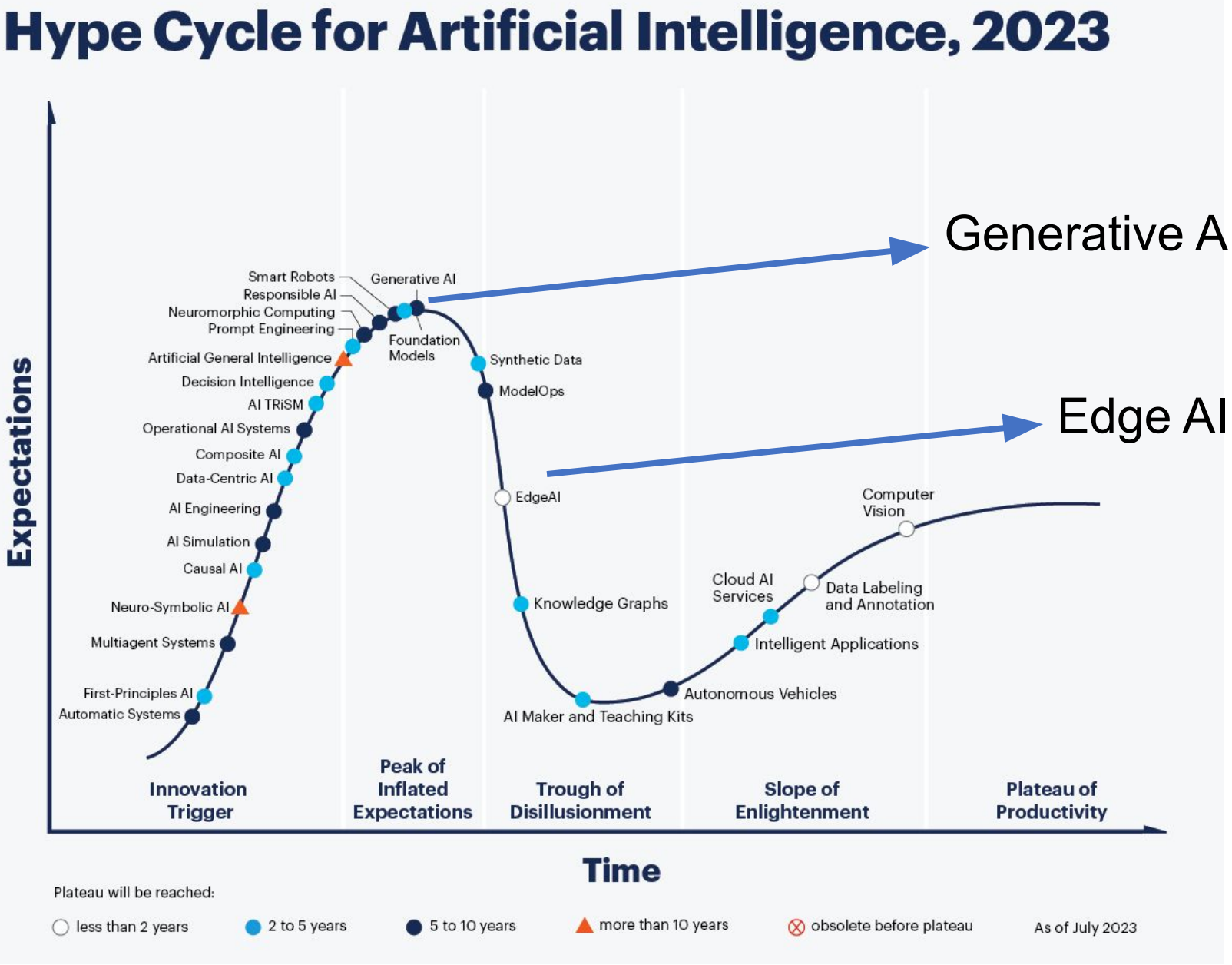
https://i.kym-cdn.com/entries/icons/original/000/044/025/shoggothhh_header.jpg

Language Models



https://www.youtube.com/watch?v=bZQun8Y4L2A&t=16s

Edge AI and Generative AI



Generative AI

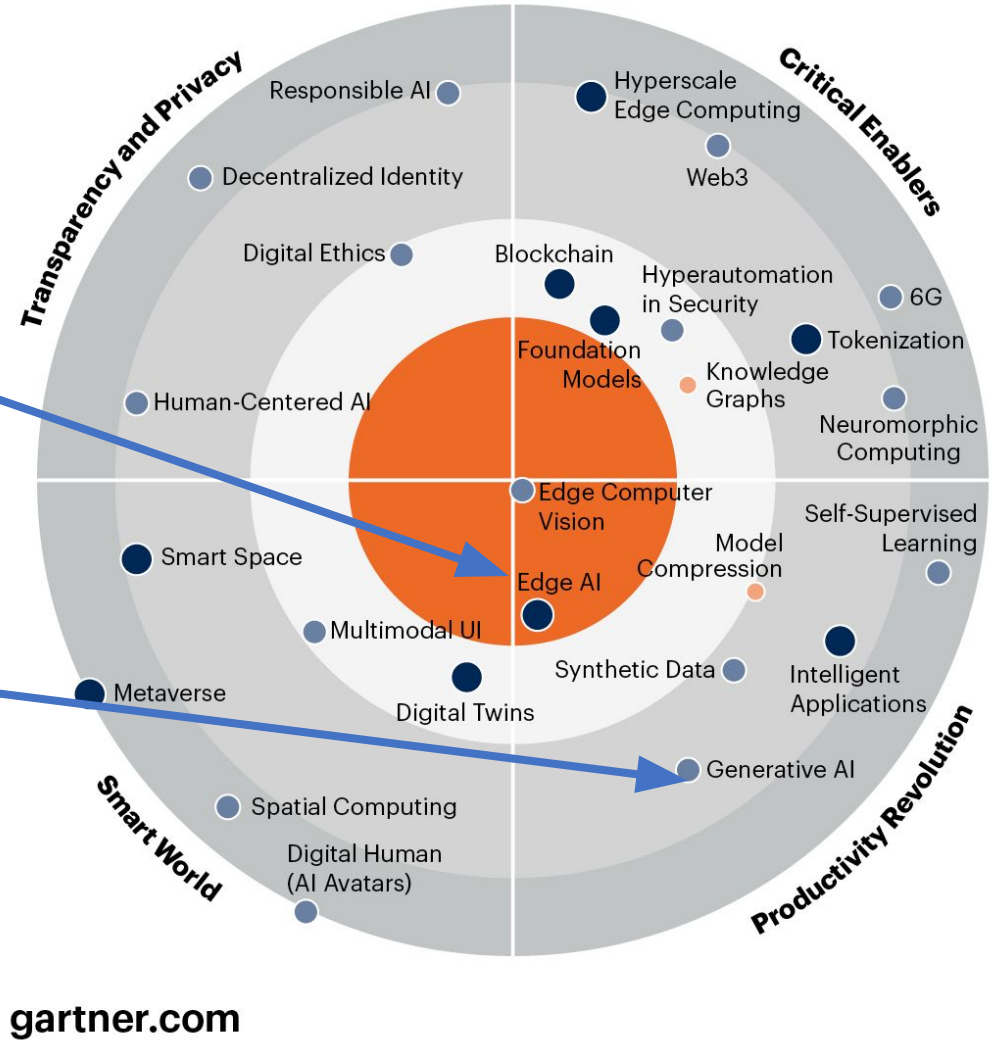
Edge AI

https://www.gartner.com/en/articles/what-s-new-in-artificial-intelligence-from-the-2023-gartner-hype-cycle

Edge AI and Generative AI

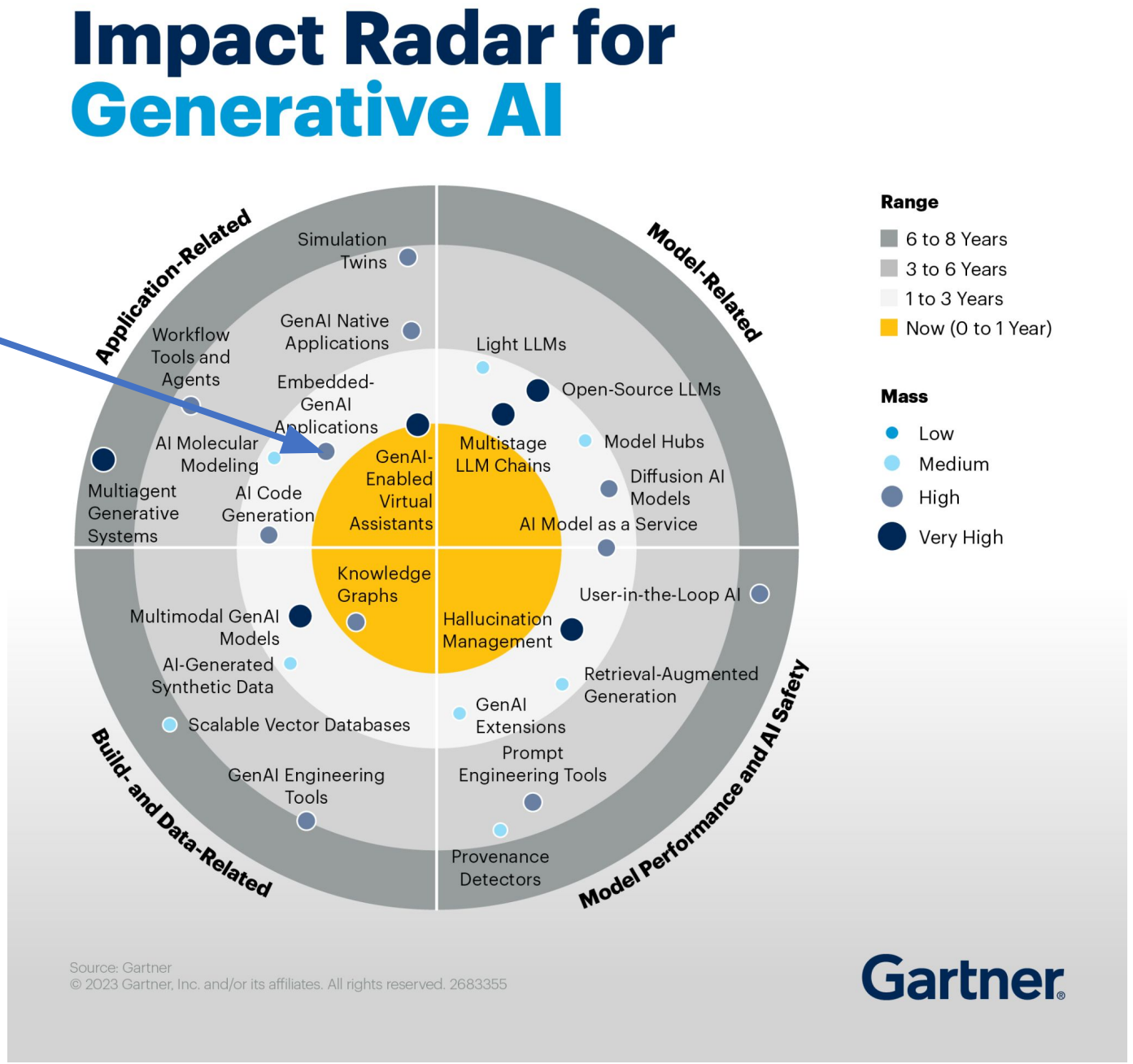
Edge AI

Generative AI



https://www.gartner.com/en/articles/4-emerging-technologies-you-need-to-know-about

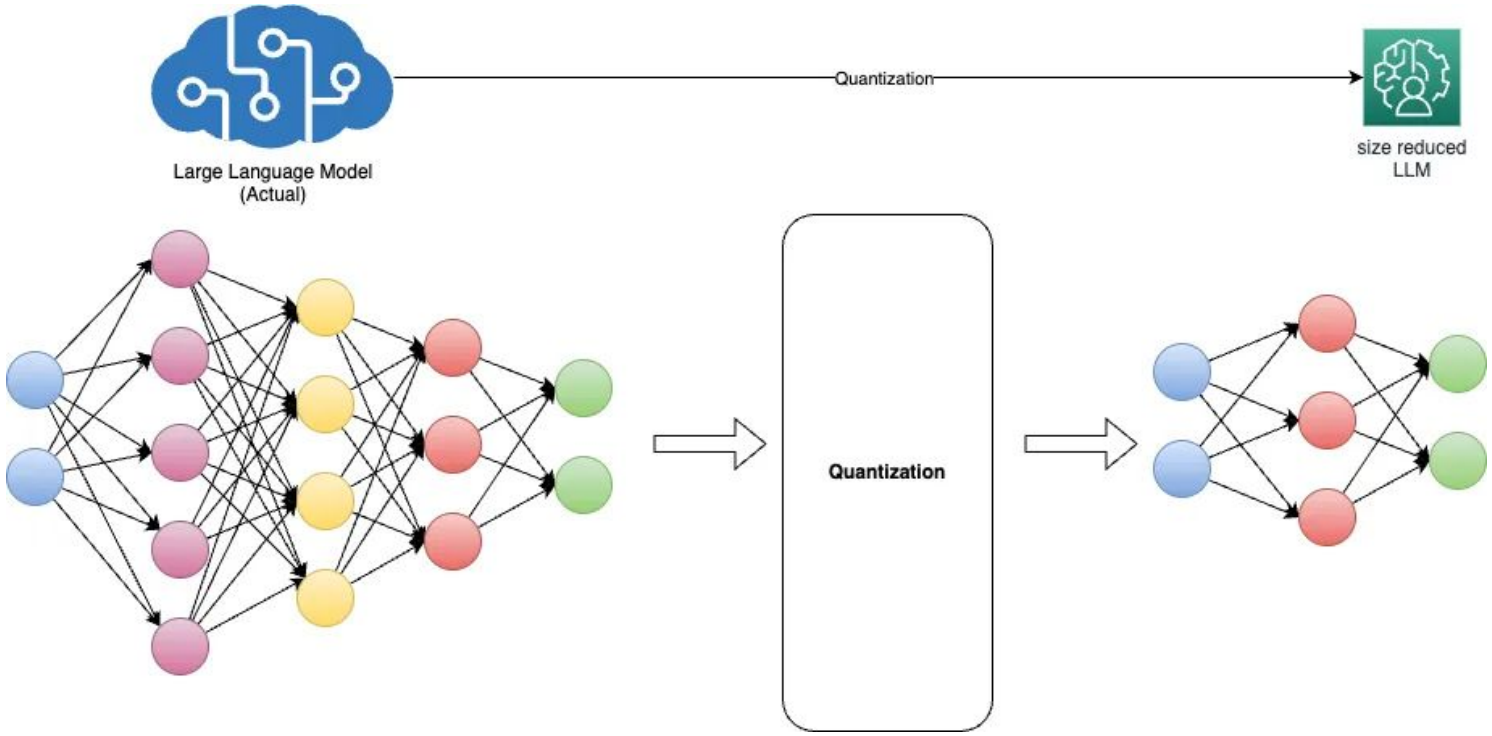
Embedded GenAI



https://www.gartner.com/en/articles/understand-and-exploit-gen-ai-with-gartner-s-new-impact-radar

Edge AI and Generative AI

- AI models optimizations
- Quantization
- Pruning
- Knowledge distillation

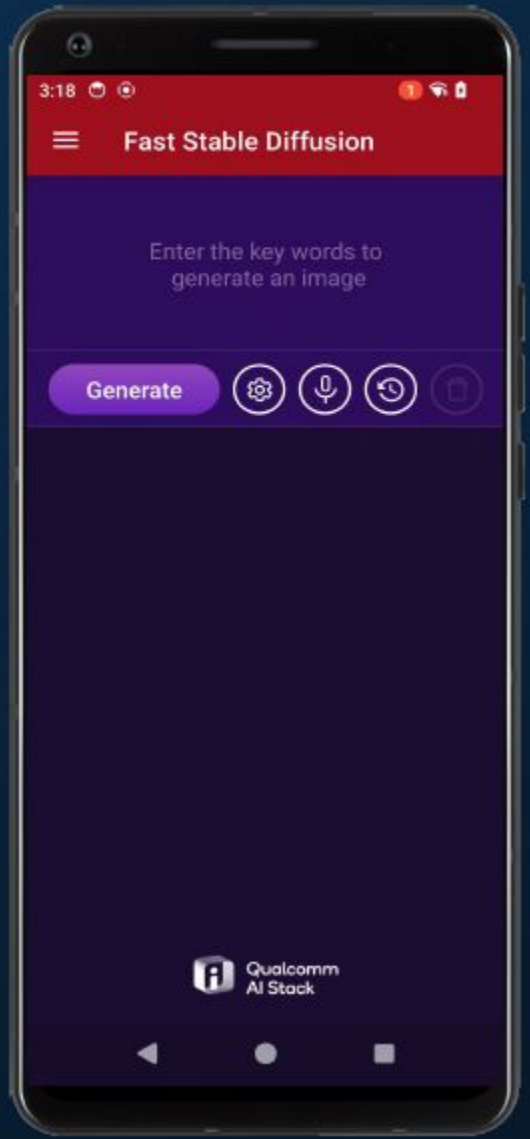


https://int8.io/local-large-language-models-beginners-guide/

https://www.linkedin.com/pulse/quantization-what-you-should-understand-want-run-llms-pavan-mantha

Edge AI and Generative AI

World's fastest AI text-to-image generative AI on a phone



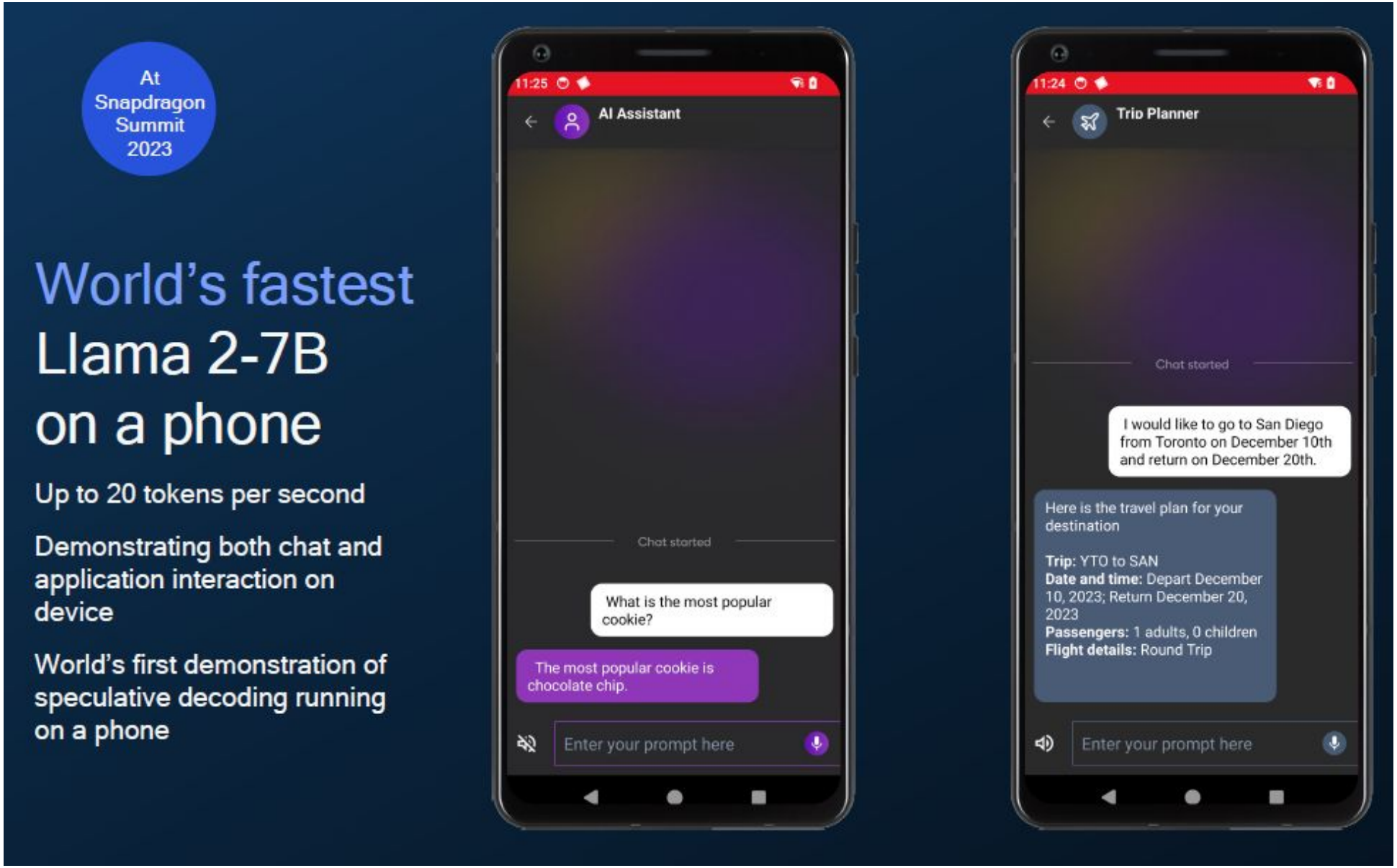
Takes less than 0.6 seconds for generating 512x512 images from text prompts

Efficient UNet architecture, guidance conditioning, and step distillation

Full-stack AI optimization to achieve this improvement

Generative AI at the edge. Joseph Soriaga Qualcomm Technologies, Inc. Nov. 8 2023

Edge AI and Generative AI



Generative AI at the edge. Joseph Soriaga Qualcomm Technologies, Inc. Nov. 8 2023

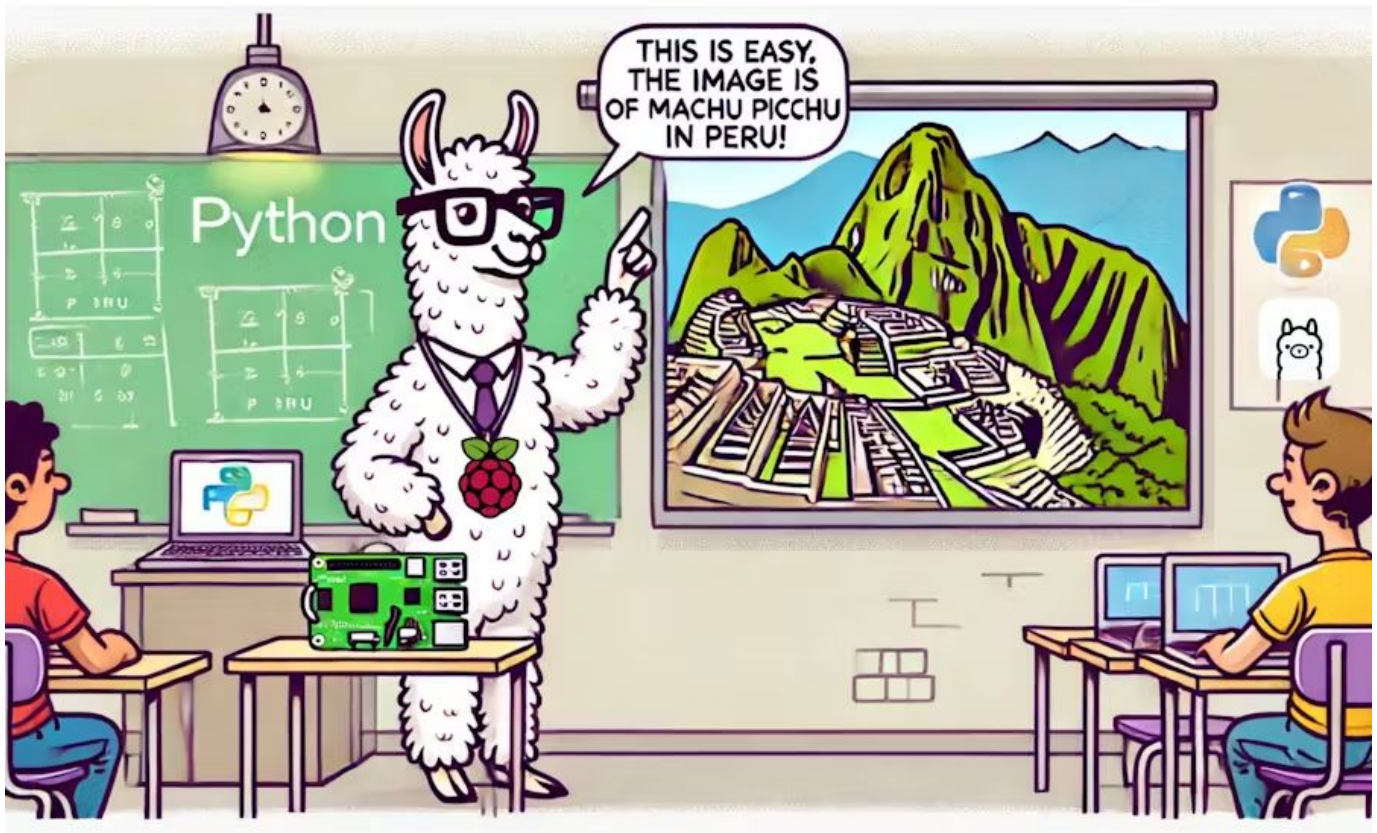
Edge AI and Generative AI

llama4micro

A "large" language model running on a microcontroller.

https://github.com/maxbbraun/llama4micro

Edge AI and Generative AI



https://www.hackster.io/mjrobot/running-large-language-models-on-raspberry-pi-at-the-edge-63bb11

Workshop on
TinyML for
Sustainable Development

Thanks

Prof. Jesús Alfonso López
jalopez@uao.edu.co
https://www.linkedin.com/in/jesus-alfonso-lópez-sotelo-76100718/
Universidad Autónoma de Occidente