

# Language Models Introduction



Prof. Jesús Alfonso López  
jalopez@uao.edu.co

Universidad Autónoma de  
Occidente

Workshop on  
TinyML for  
Sustainable Development

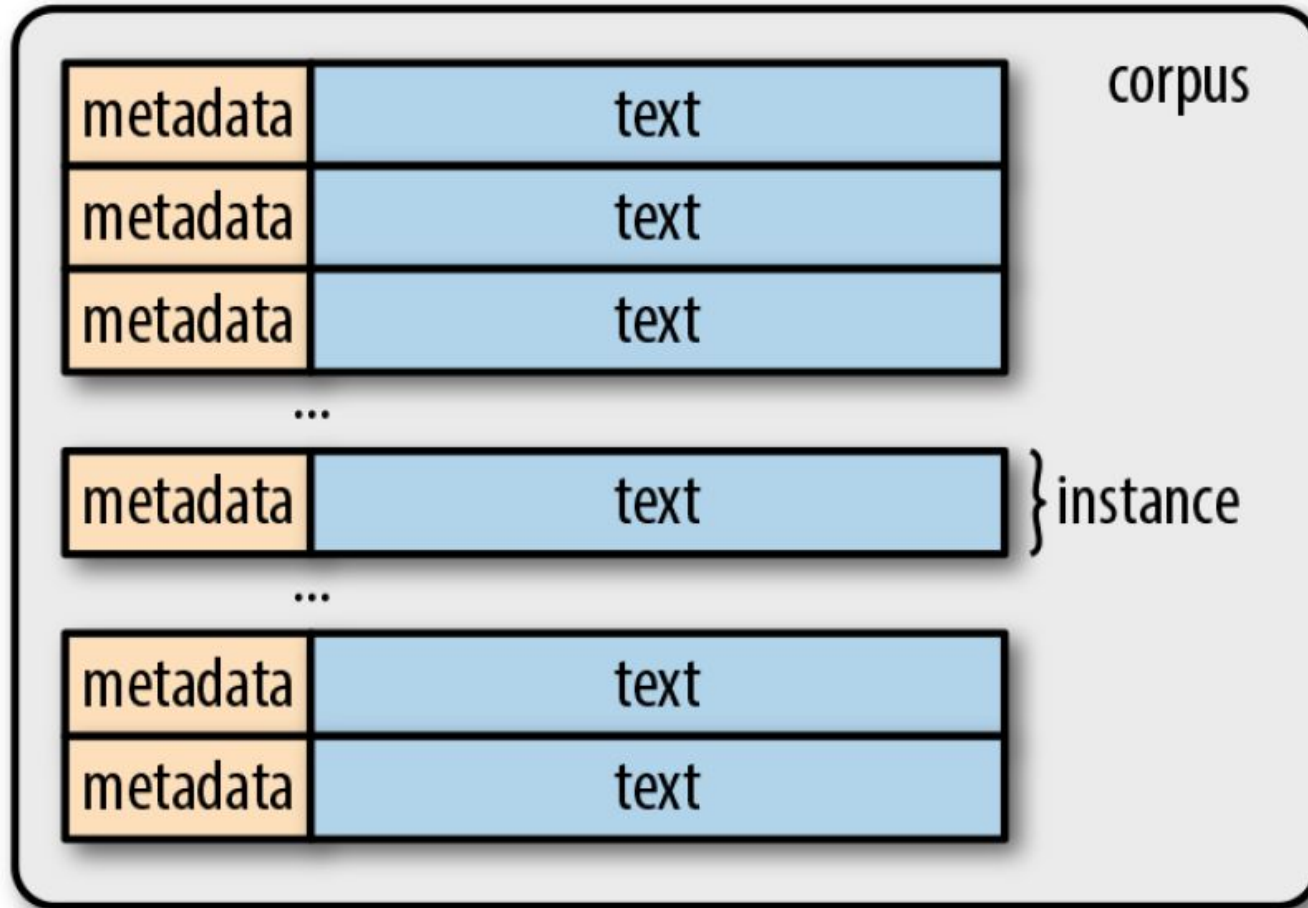


The Abdus Salam  
International Centre  
for Theoretical Physics



# 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

GPT-3.5 & GPT-4 GPT-3 (Legacy)

Probemos el tokenizado de GPT-4 con una frase en español.  
Recordemos que un token puede ser en algunos casos igual a una palabra o a una parte o fracción de palabra|

Clear

Show example

Tokens	Characters
39	164

Probemos el tokenizado de GPT-4 con una frase en español.  
Recordemos que un token puede ser en algunos casos igual a una palabra o a una parte o fracción de palabra

Text

Token IDs

<https://platform.openai.com/tokenizer>

# Language Models

## Token



Simon Willison



By Simon Willison · Edited Jun 8, 2023 · 16 forks · 40 Likes

### GPT token encoder and decoder

For more information on this tool, read [Understanding GPT tokenizers](#)

Enter text to tokenize it:

16281 2420 318 994

4 tokens

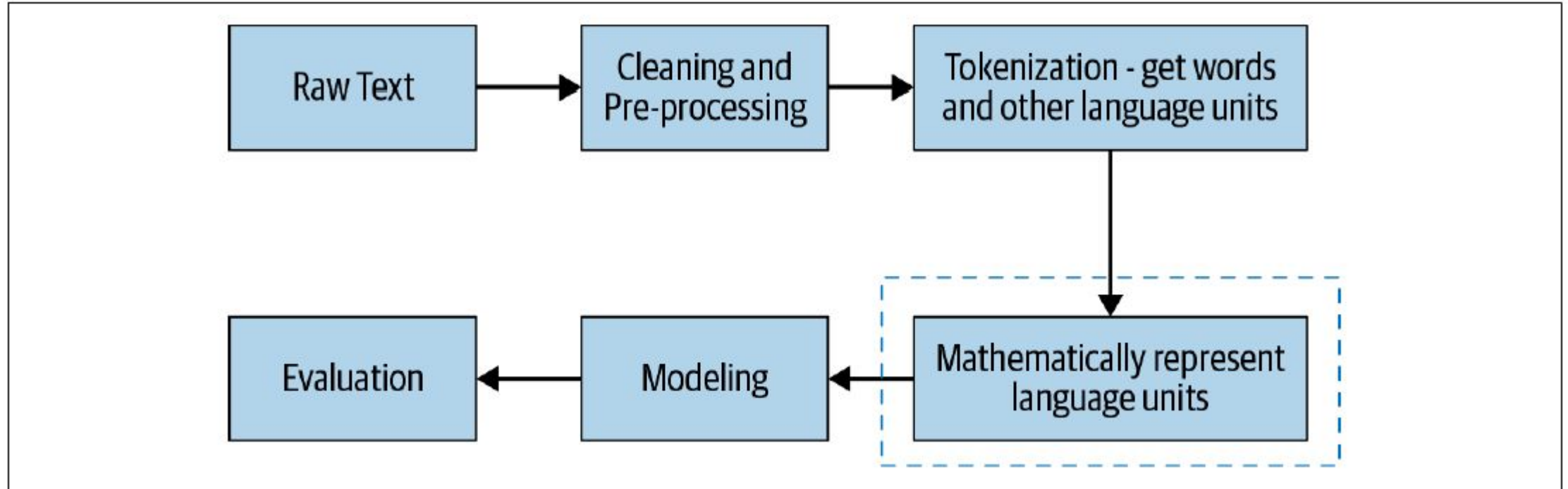
+ > +

Example	text	is	here
16281	2420	318	994

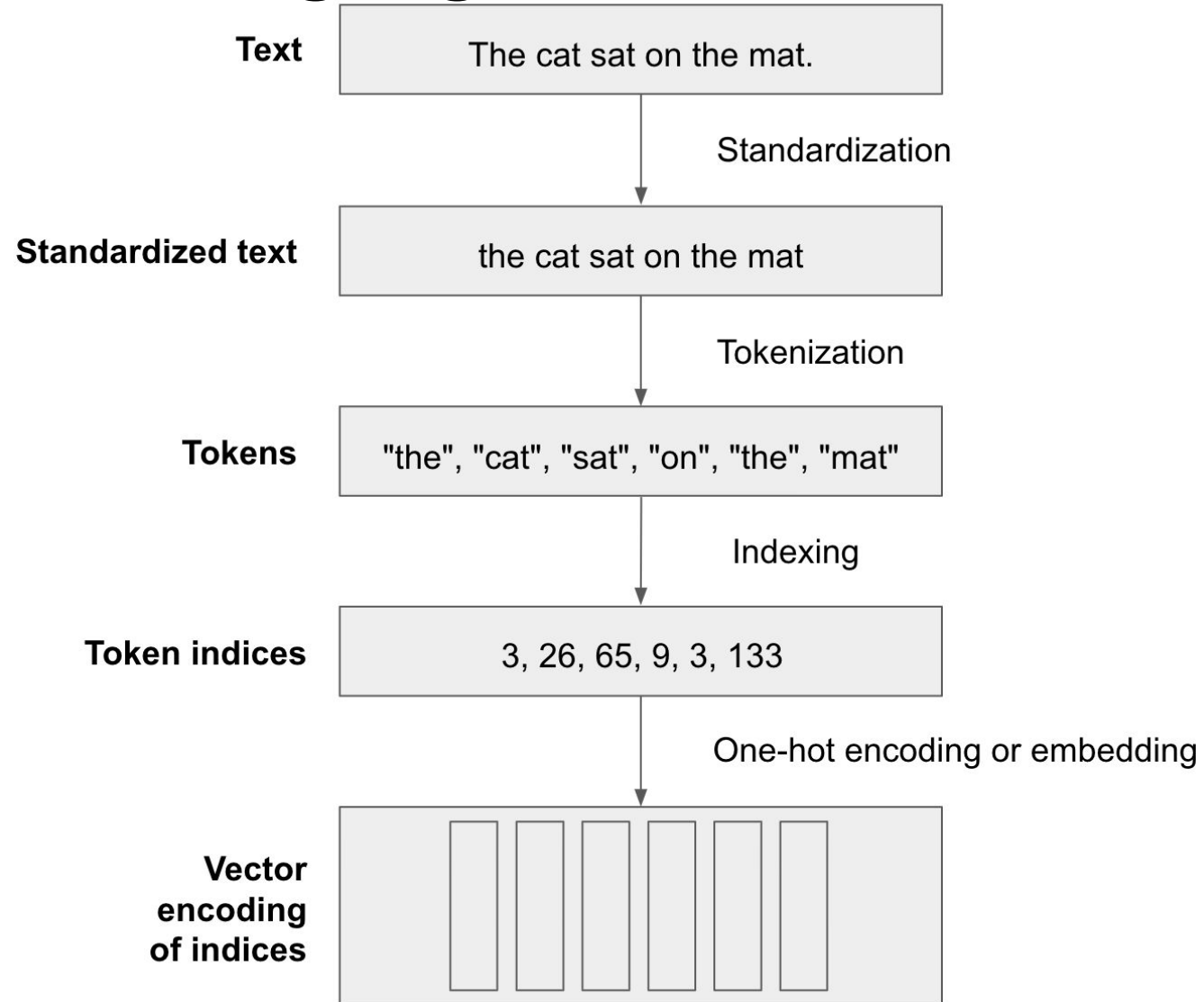
Or convert tokens to text:

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

# Language Models



# Language Models

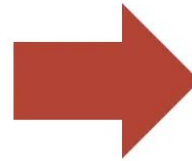


# Language Models

Word embedding = From tokens to vectors

One hot encoding

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



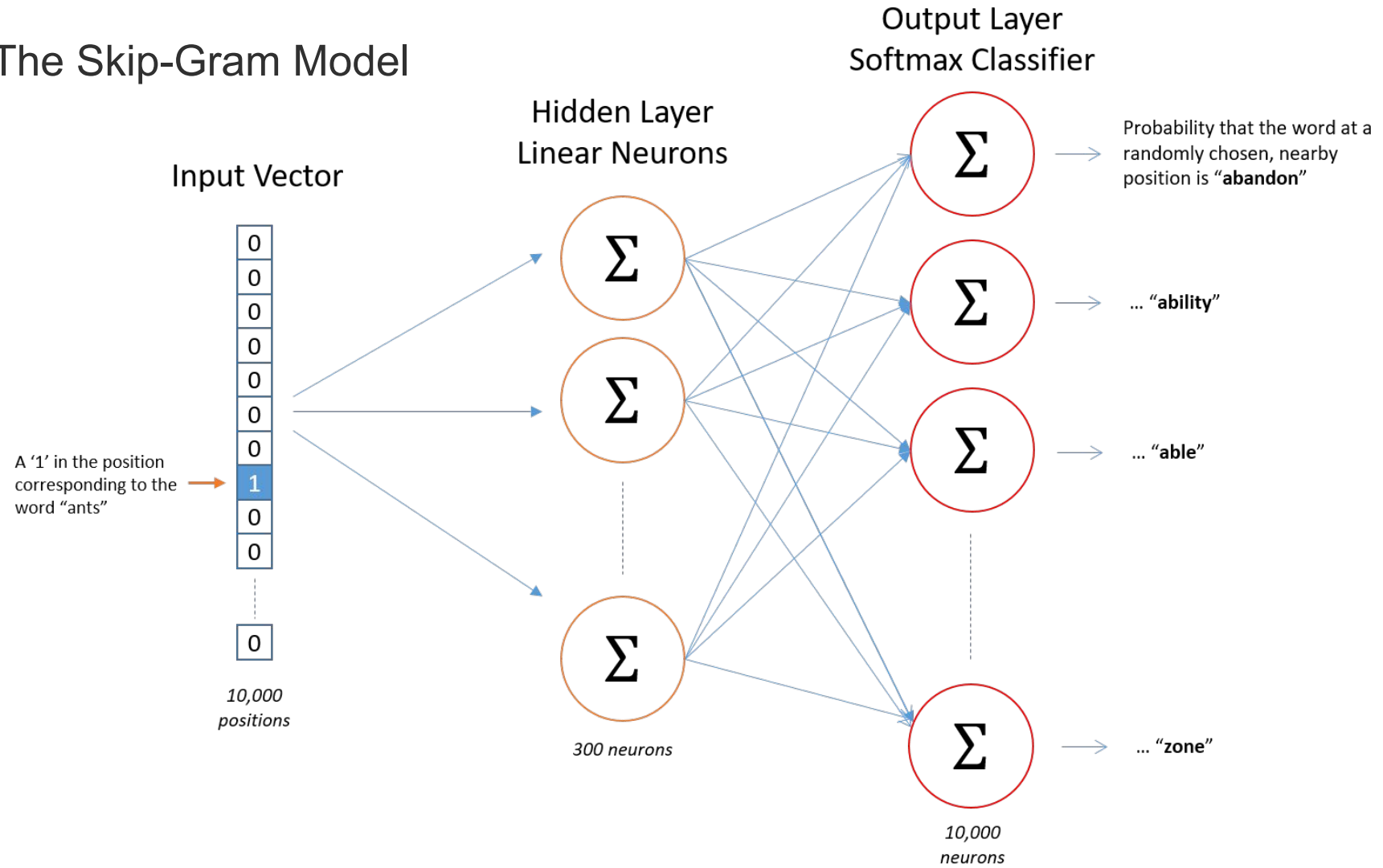
	1	2	3	4	5	6	7	8	9
man	1	0	0	0	0	0	0	0	0
woman	0	1	0	0	0	0	0	0	0
boy	0	0	1	0	0	0	0	0	0
girl	0	0	0	1	0	0	0	0	0
prince	0	0	0	0	1	0	0	0	0
princess	0	0	0	0	0	1	0	0	0
queen	0	0	0	0	0	0	1	0	0
king	0	0	0	0	0	0	0	1	0
monarch	0	0	0	0	0	0	0	0	1

Each word gets  
a 1x9 vector  
representation

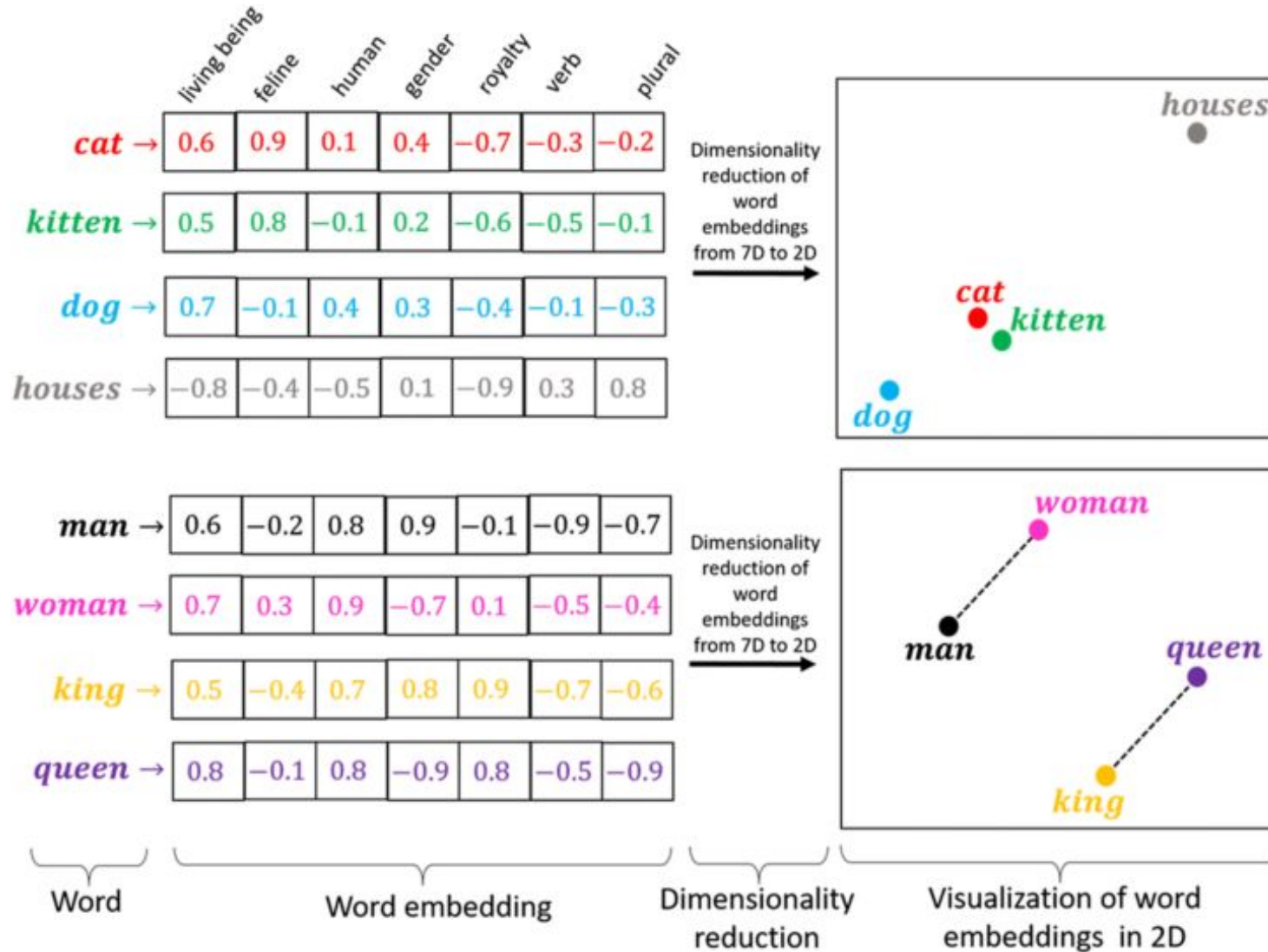


# Language Models

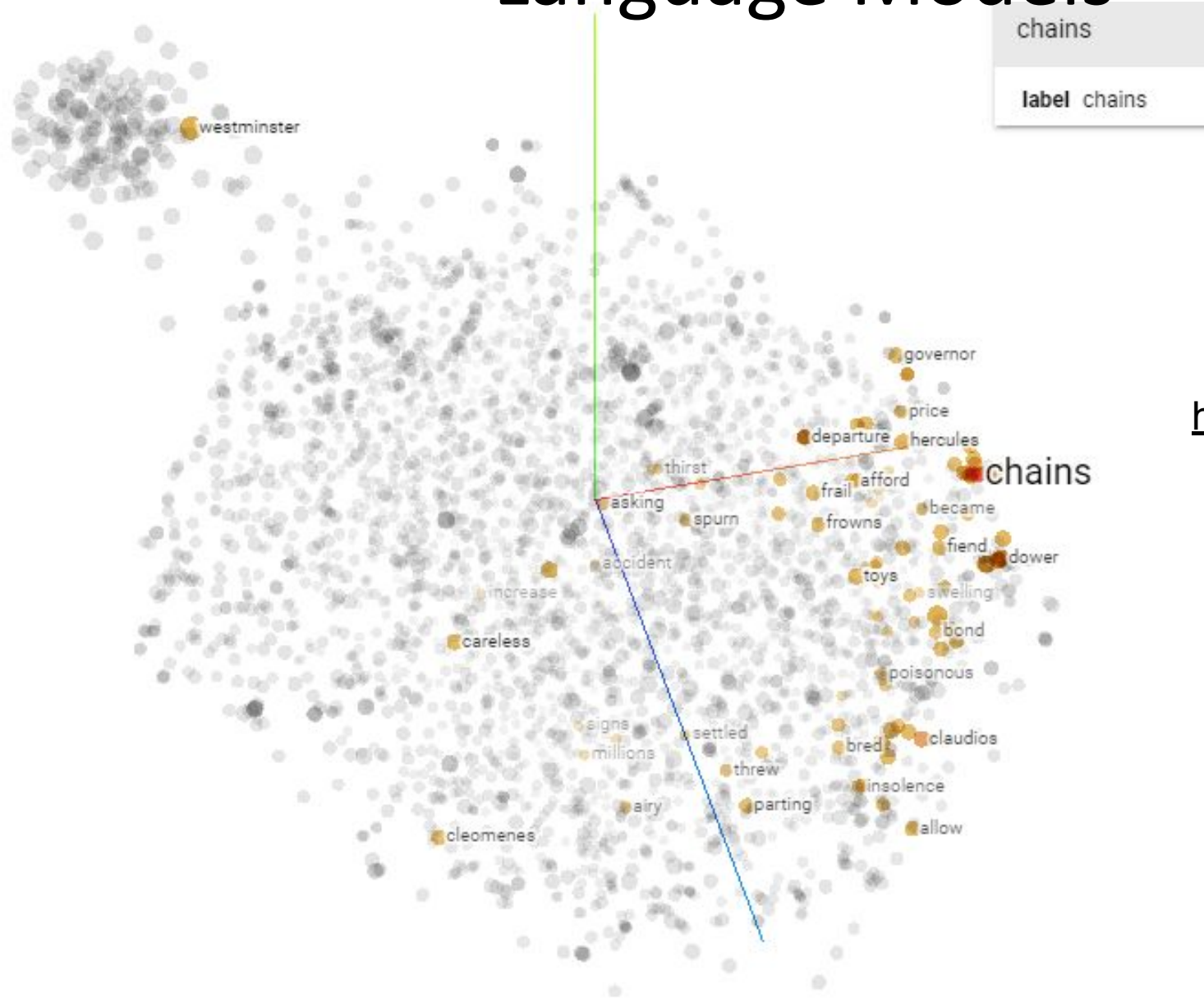
## The Skip-Gram Model



# Language Models

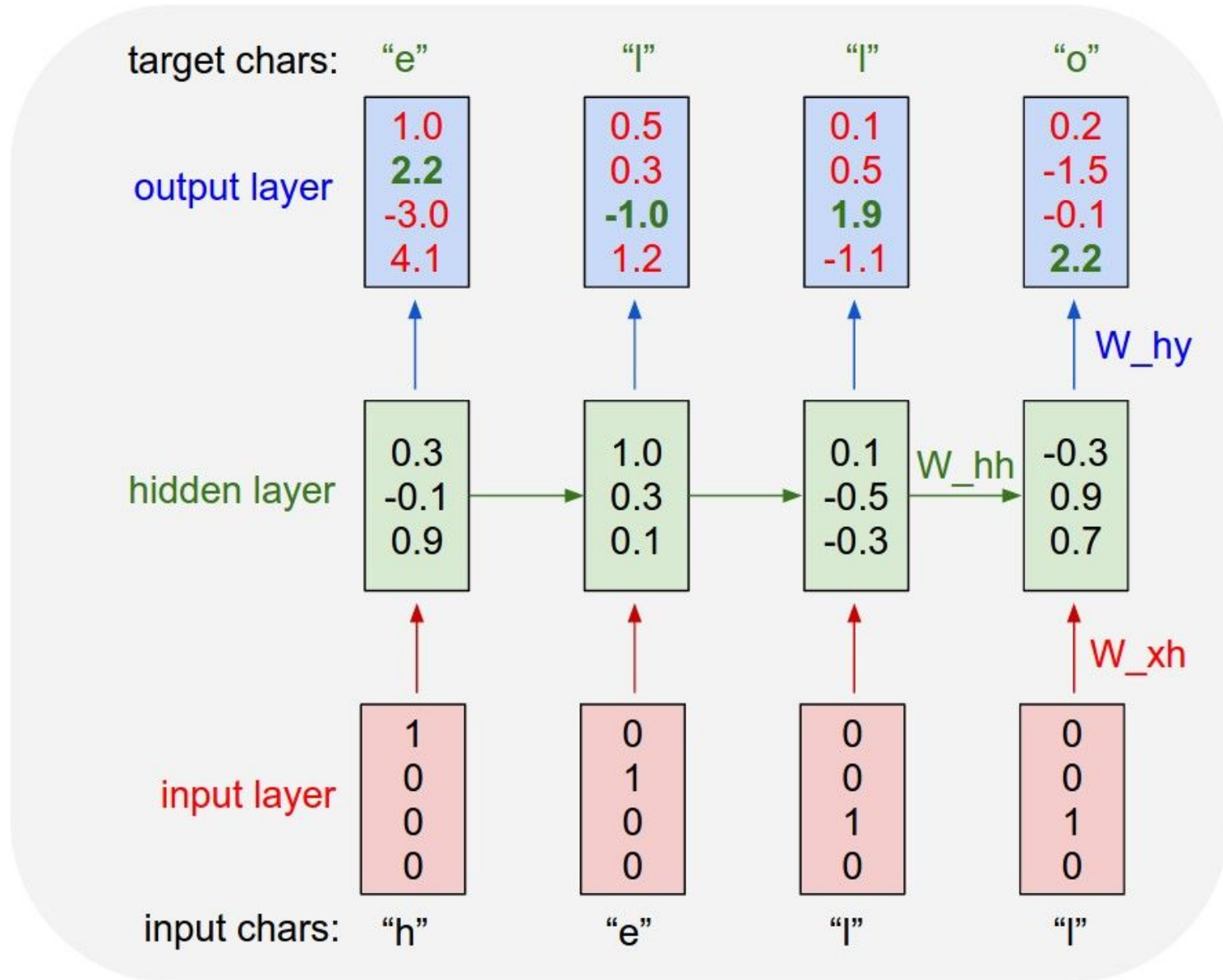


# Language Models



<https://projector.tensorflow.org/>

# Language Models



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

# Language Models

RNNs: recurrence to model sequence dependencies

## Limitations of RNNs



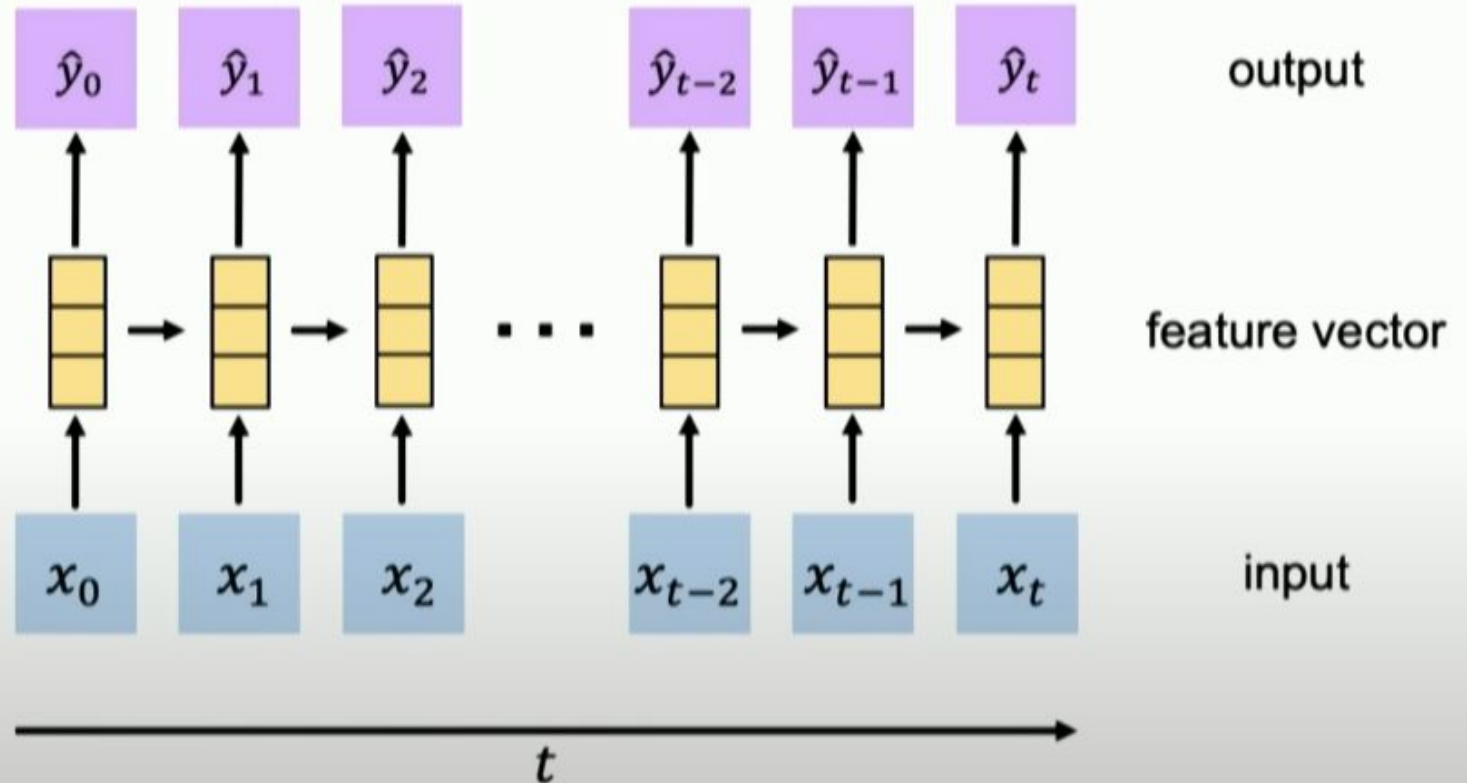
Encoding bottleneck



Slow, no parallelization



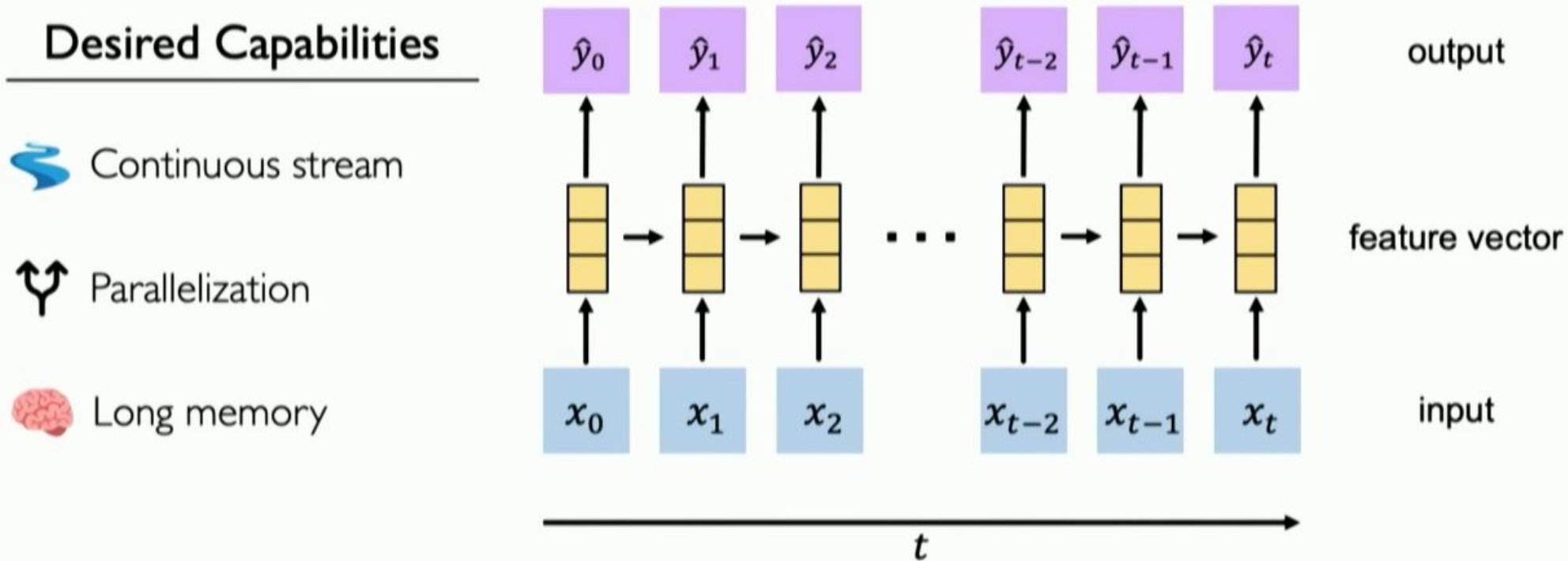
Not long memory



[https://www.youtube.com/watch?v=ySEx\\_Bqxxvvo](https://www.youtube.com/watch?v=ySEx_Bqxxvvo)

# Language Models

## Goal of Sequence Modeling



[https://www.youtube.com/watch?v=ySEx\\_Bqyvvo](https://www.youtube.com/watch?v=ySEx_Bqyvvo)

# 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

**Lukasz Kaiser\***  
Google Brain  
lukaszkaizer@google.com

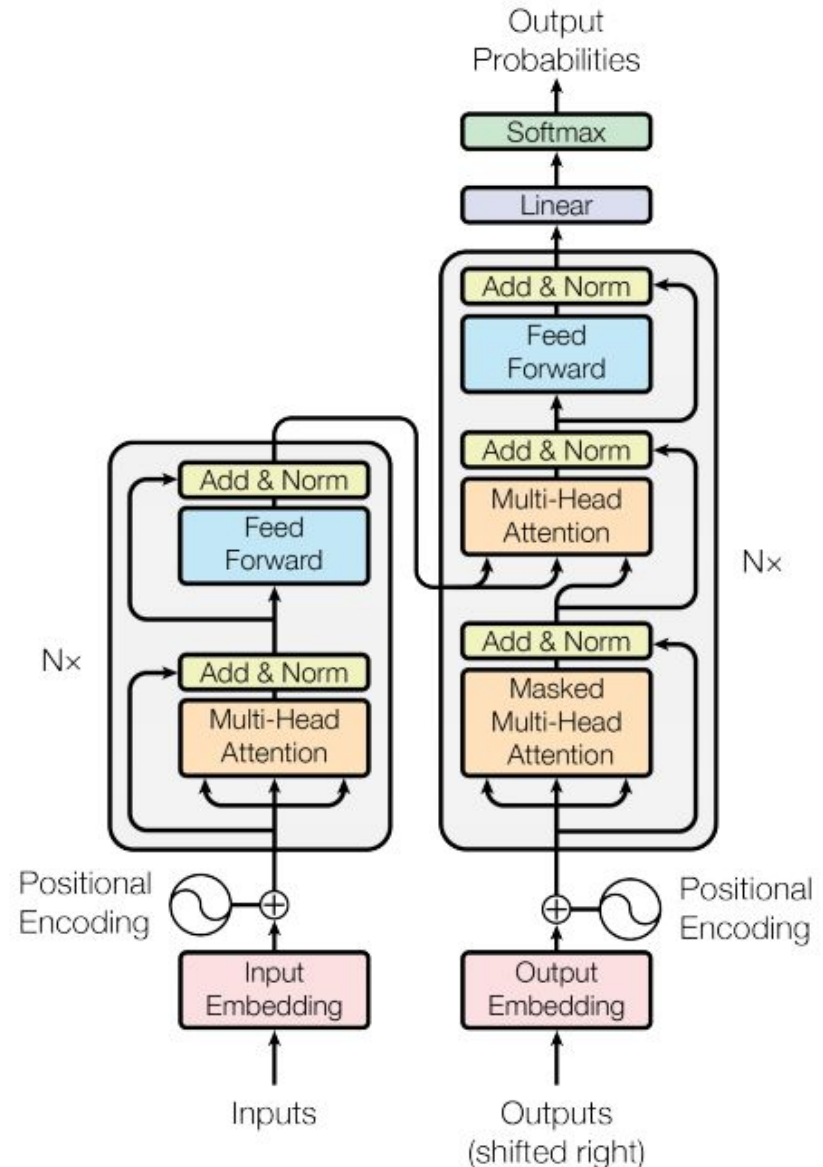
**Illia Polosukhin\* ‡**  
illia.polosukhin@gmail.com

### 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>

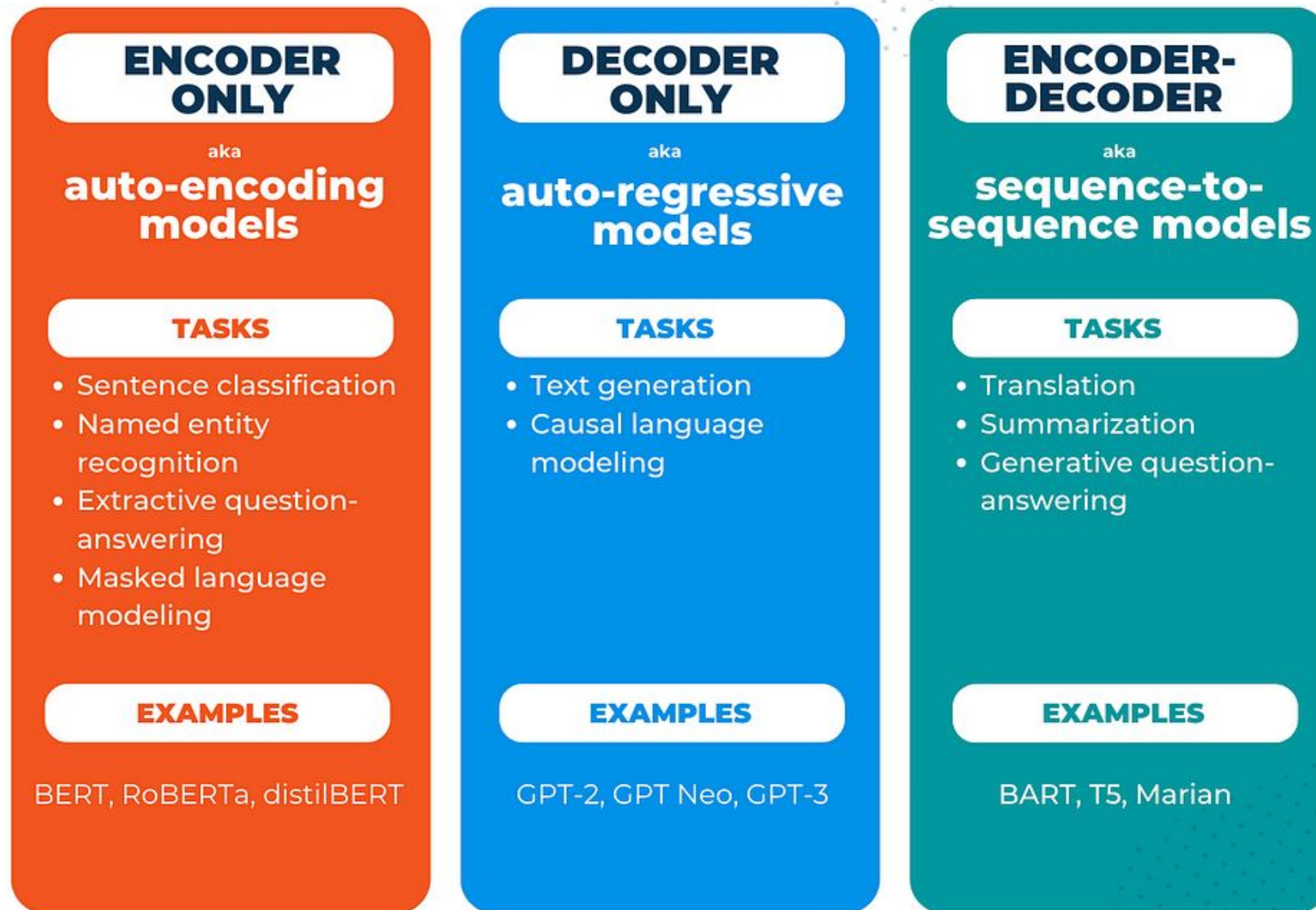
<http://jalammar.github.io/illustrated-transformer/>

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

The screenshot displays a user interface for exploring GPT models. At the top, there are navigation buttons for 'GPT-2 (small)', 'nano-gpt', 'GPT-2 (XL)', and 'GPT-3'. The 'nano-gpt' model is selected, with a box indicating it has 85,584 parameters. On the left, a diagram shows the LLM architecture. It starts with an input 'How to predict text' (tokens: 2437, 284, 4331; words: 2420, 16326, 2456). This input goes through a 'tok embed' and 'pos embed' block, which are summed. The result enters 'transformer i', which consists of a 'layer norm', 'multi-head, causal self-attention', another 'layer norm', a 'feed forward' block, a third 'layer norm', a 'linear' block, and a 'softmax' block. A 'Table of Contents' on the right lists: Intro, Introduction, Preliminaries, Components, Embedding, Layer Norm, Self Attention, Projection, MLP, Transformer, Softmax, and Output. Below the diagram, a text box says: 'Welcome to the walkthrough of the GPT large language model! Here we'll explore the model nano-gpt, with a mere 85,000 parameters. Its goal is a simple one: take a sequence of six letters: GRABBC'. At the bottom, there are 'Continue' and 'Skip' buttons. On the right side of the interface, there is a 3D visualization of the model's internal structure, showing multiple layers of transformer blocks connected by green lines.



# Language Models



# Language Models

## Grandes Modelos de lenguaje (Large Language Model)

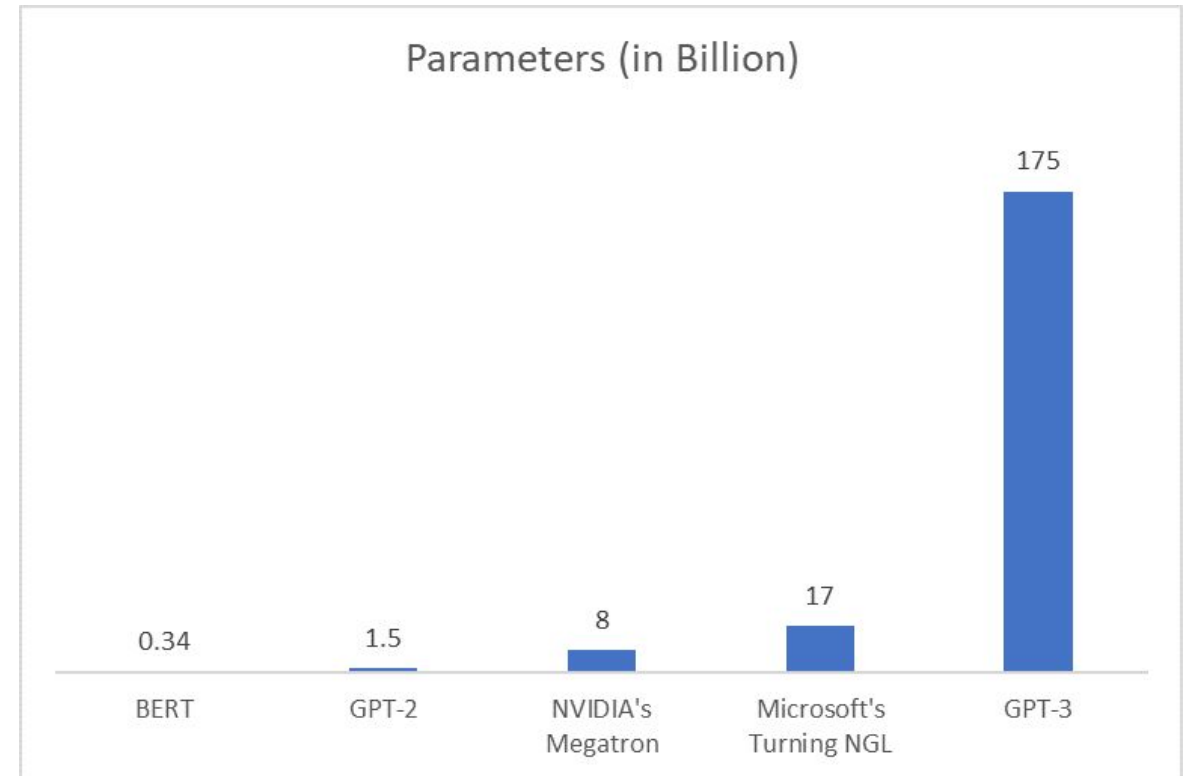
Year	Model	# of Parameters	Dataset Size
2019	BERT [39]	3.4E+08	16GB
2019	DistilBERT [113]	6.60E+07	16GB
2019	ALBERT [70]	2.23E+08	16GB
2019	XLNet (Large) [150]	3.40E+08	126GB
2020	ERNIE-GEN (Large) [145]	3.40E+08	16GB
2019	RoBERTa (Large) [74]	3.55E+08	161GB
2019	MegatronLM [122]	8.30E+09	174GB
2020	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)

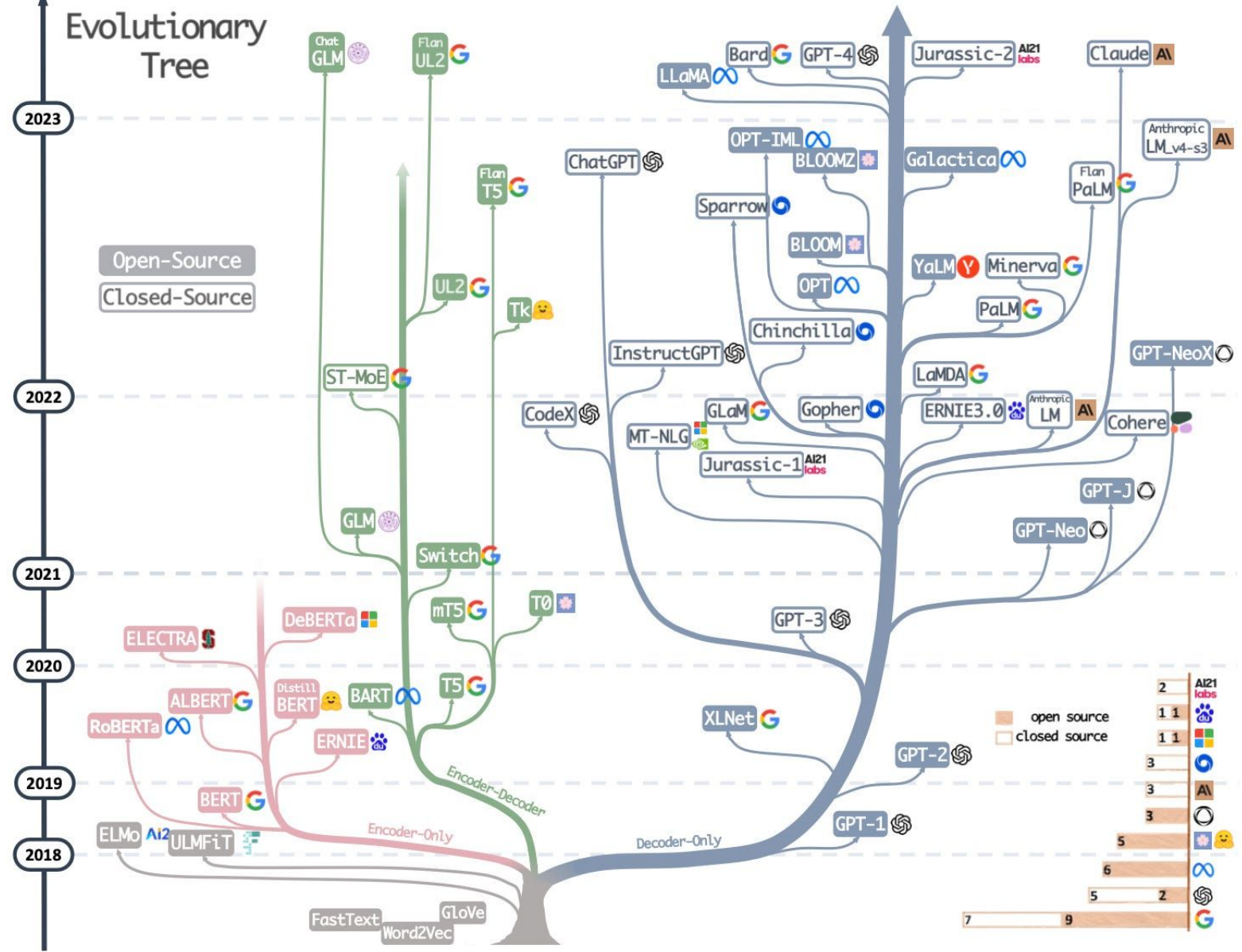
5.30 E+11



[http://faculty.washington.edu/ebender/papers/Stochastic\\_Parrots.pdf](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>

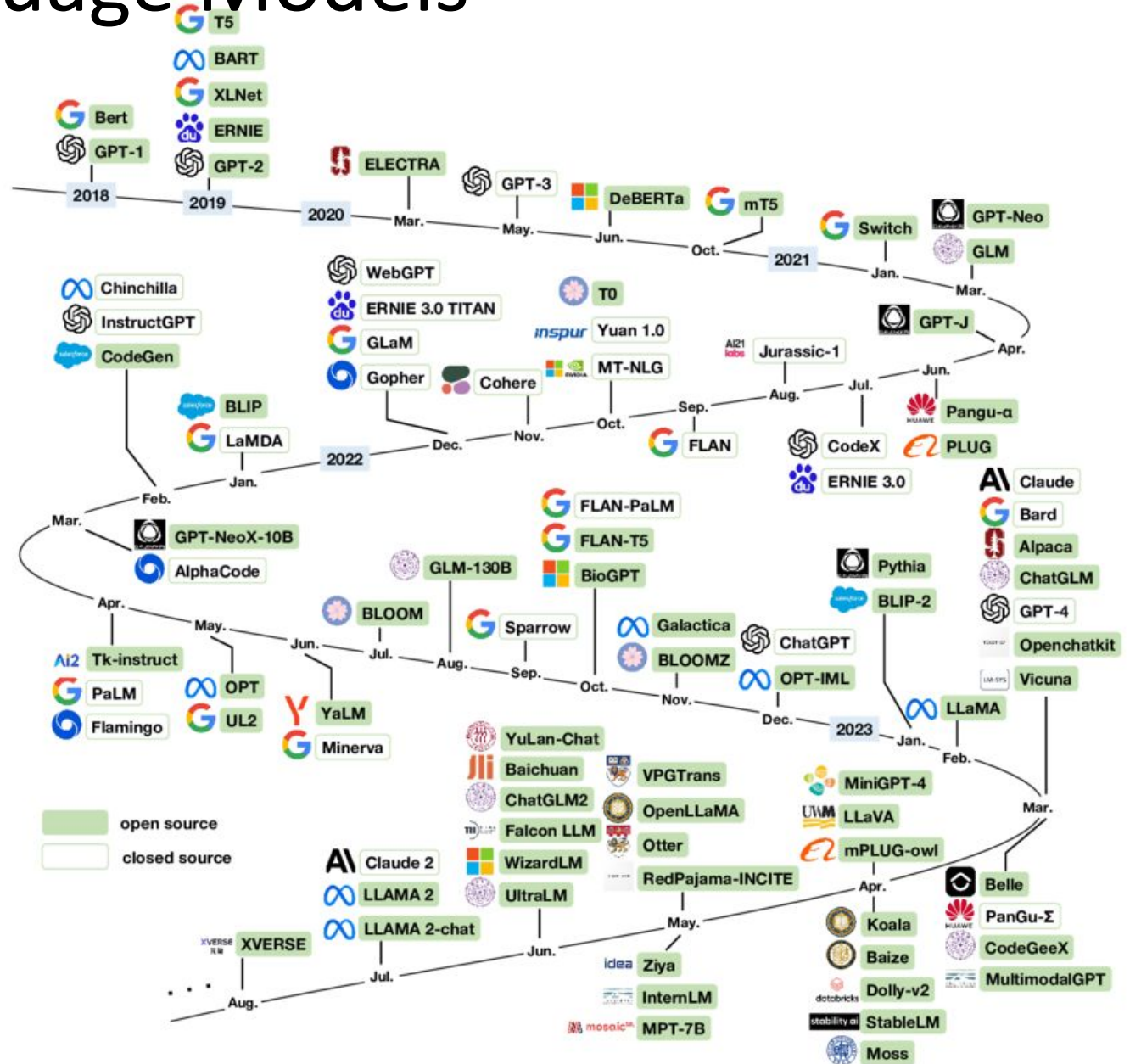
# Language Models



LLM Evolution

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

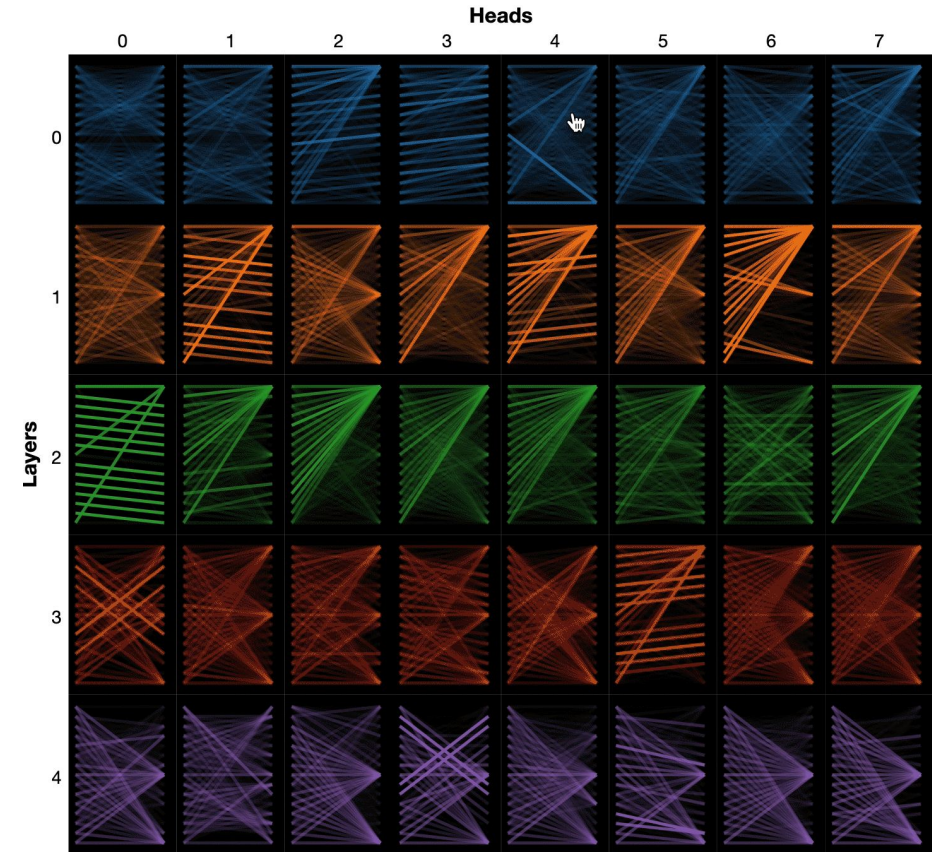
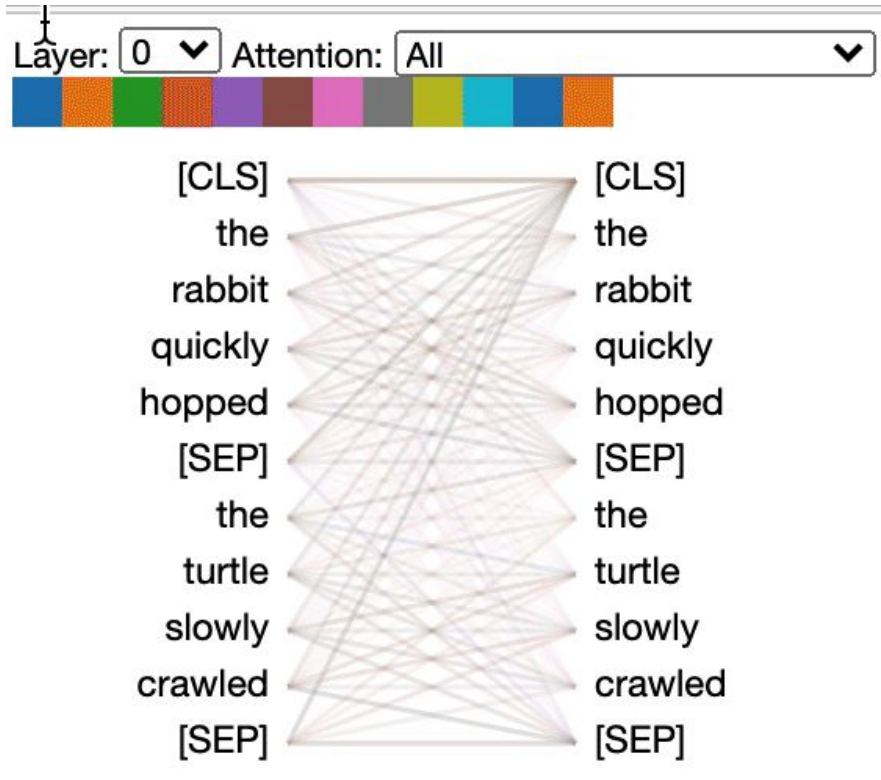
# Language Models



LLM  
Evolution

[https://www.researchgate.net/figure/A-chronological-overview-of-large-language-models-LLMs-multimodal-and-scientific\\_fig2\\_373451304](https://www.researchgate.net/figure/A-chronological-overview-of-large-language-models-LLMs-multimodal-and-scientific_fig2_373451304)

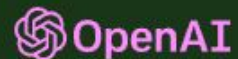
# Language Models



<https://github.com/jessevig/bertviz>

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

# Language Models



## ChatGPT: Optimizing Language Models for Dialogue

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests. ChatGPT is a sibling model to [InstructGPT](#), which is trained to follow an instruction in a prompt and provide a detailed response.

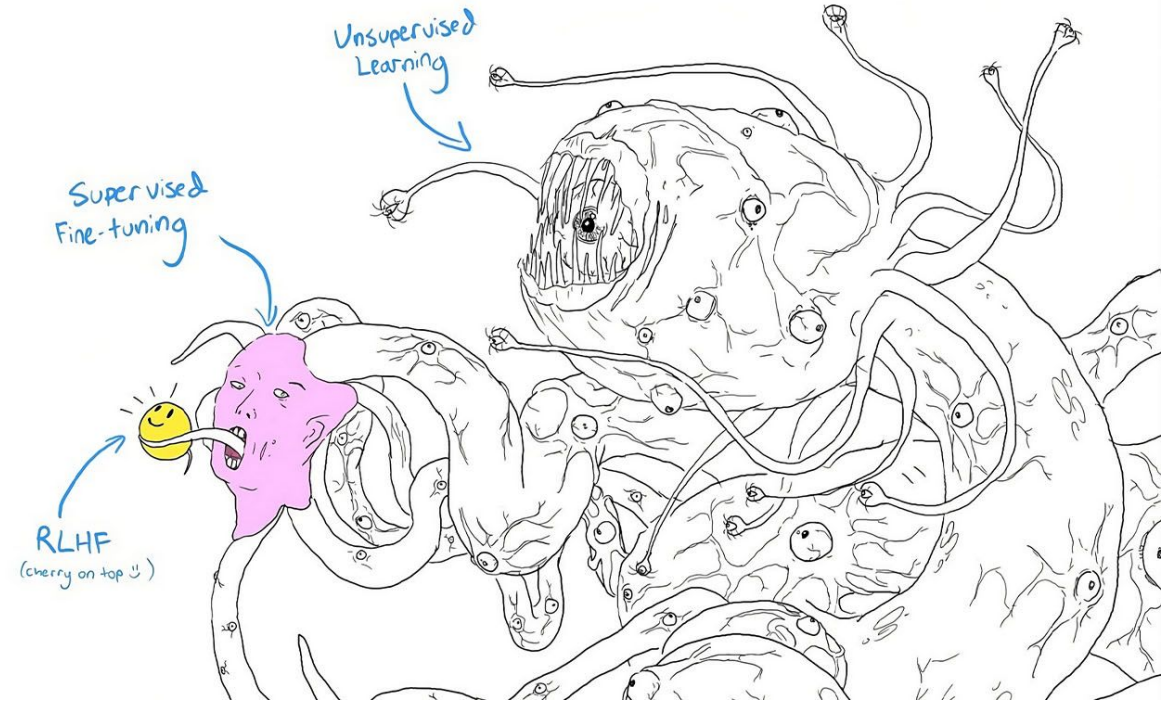
TRY CHATGPT ↗

<https://chatgpt.com/>

# Language Models

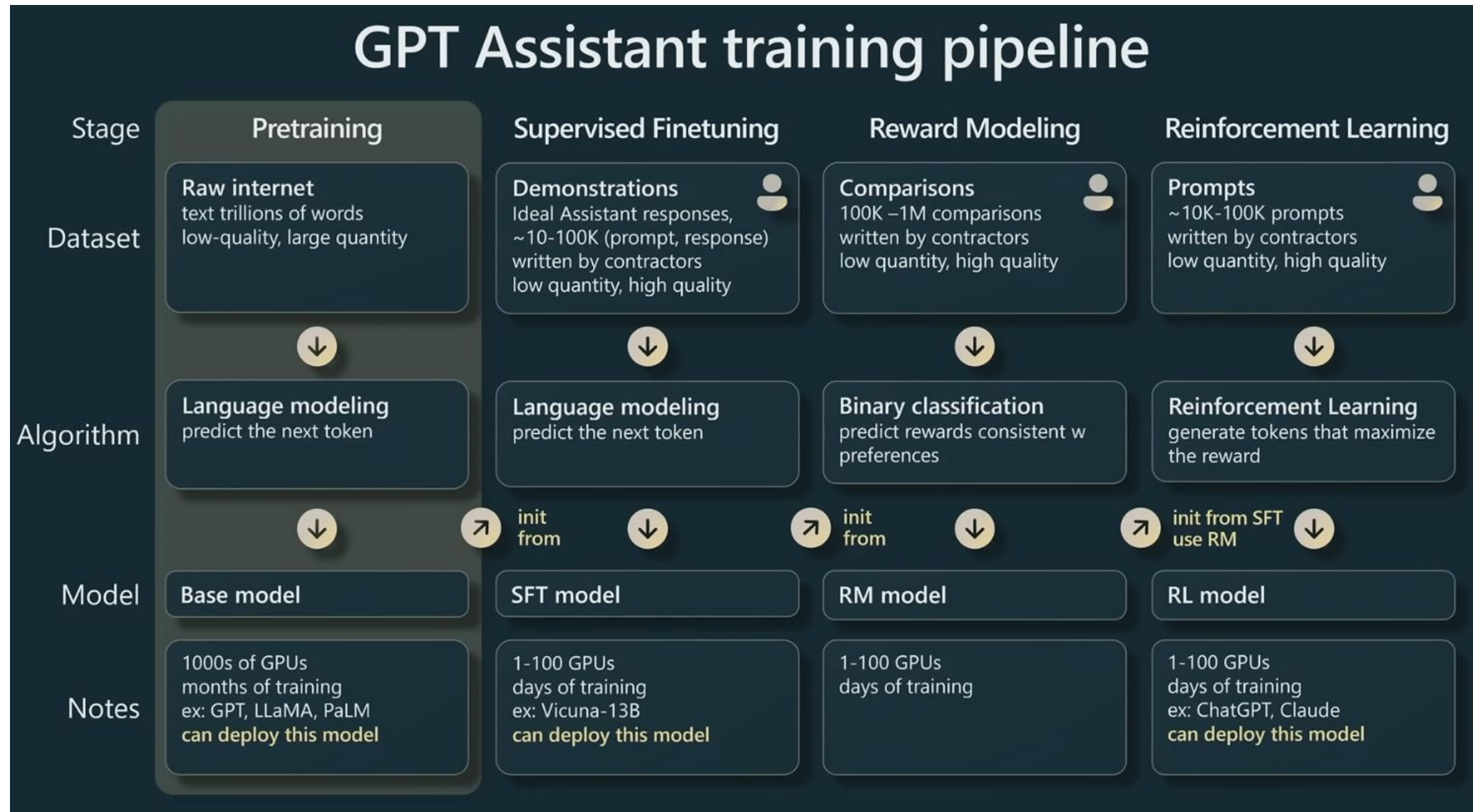
## Reinforcement learning from Human Feedback

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

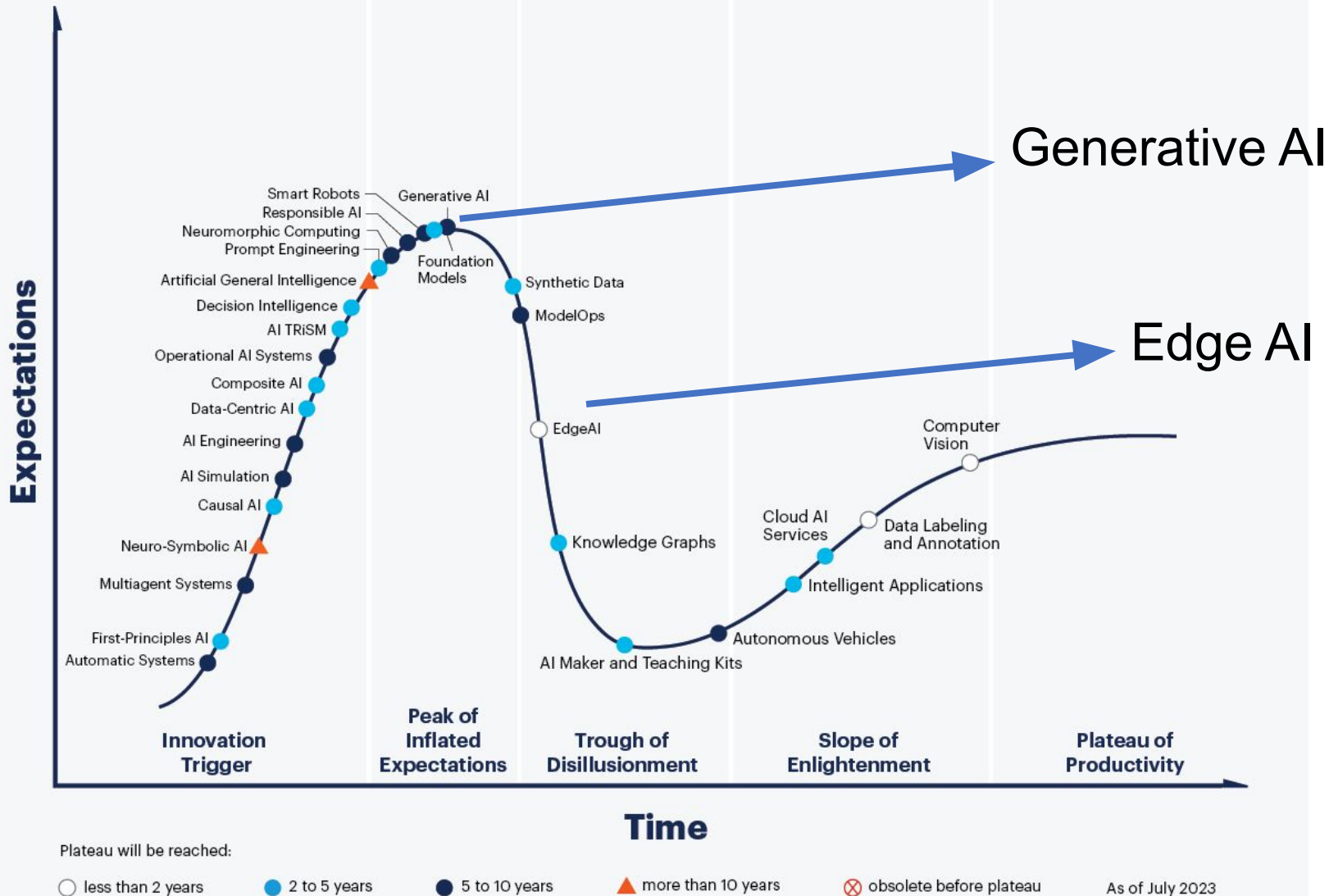
# Language Models



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



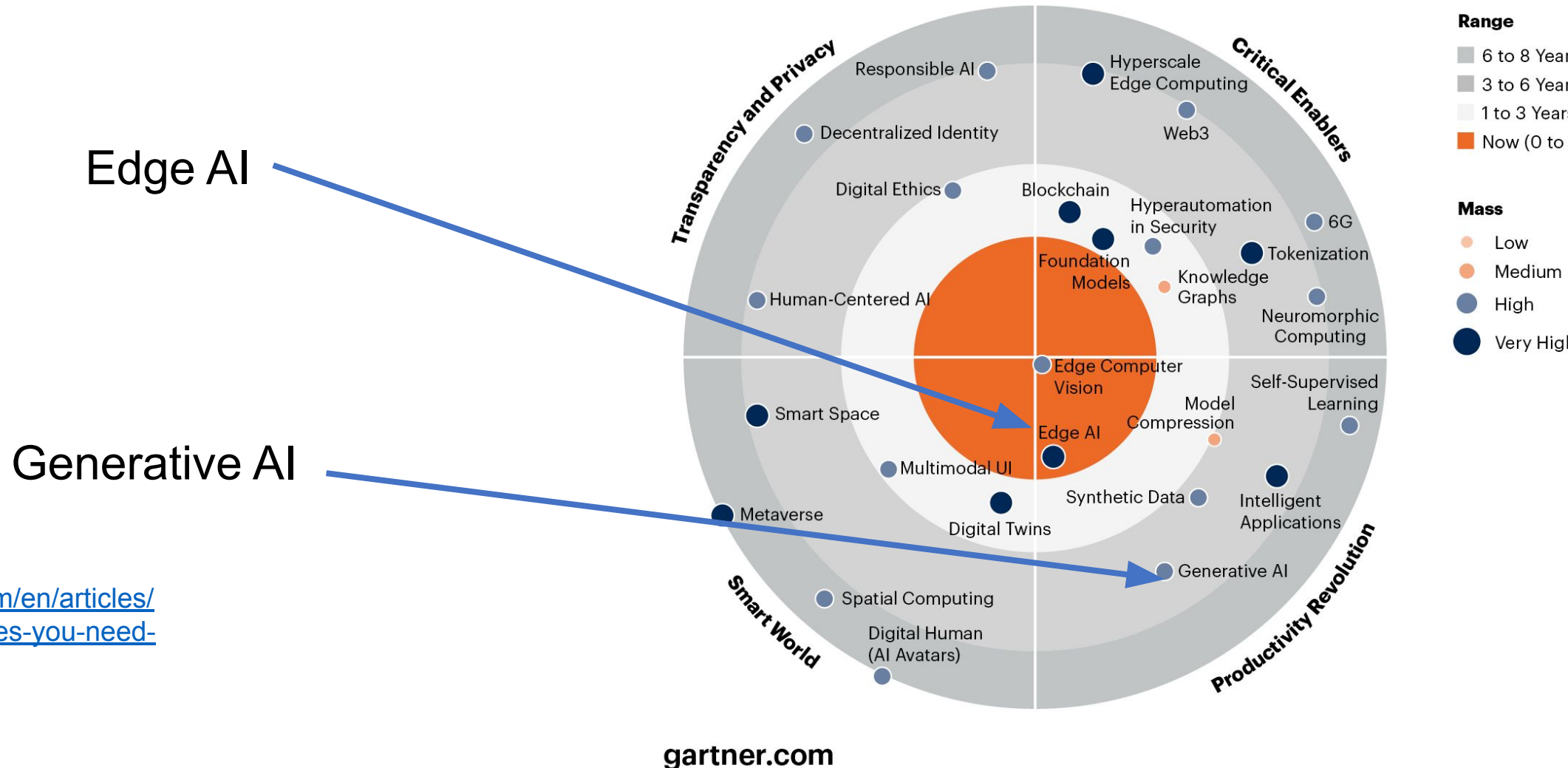
## Hype Cycle for Artificial Intelligence, 2023



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

# Edge AI and Generative AI

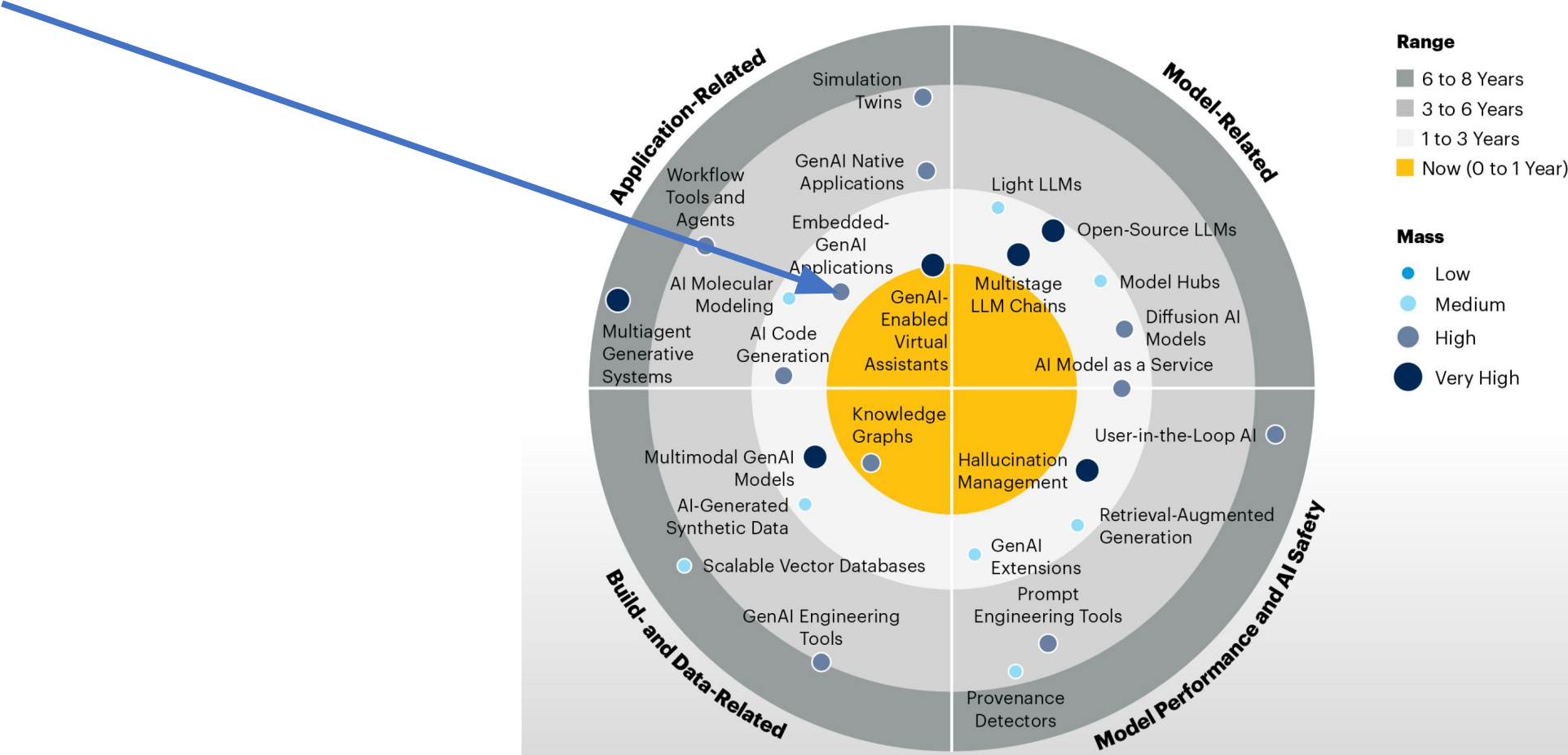
## 2023 Gartner Emerging Technologies and Trends Impact Radar



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

# Impact Radar for Generative AI

Embedded GenAI



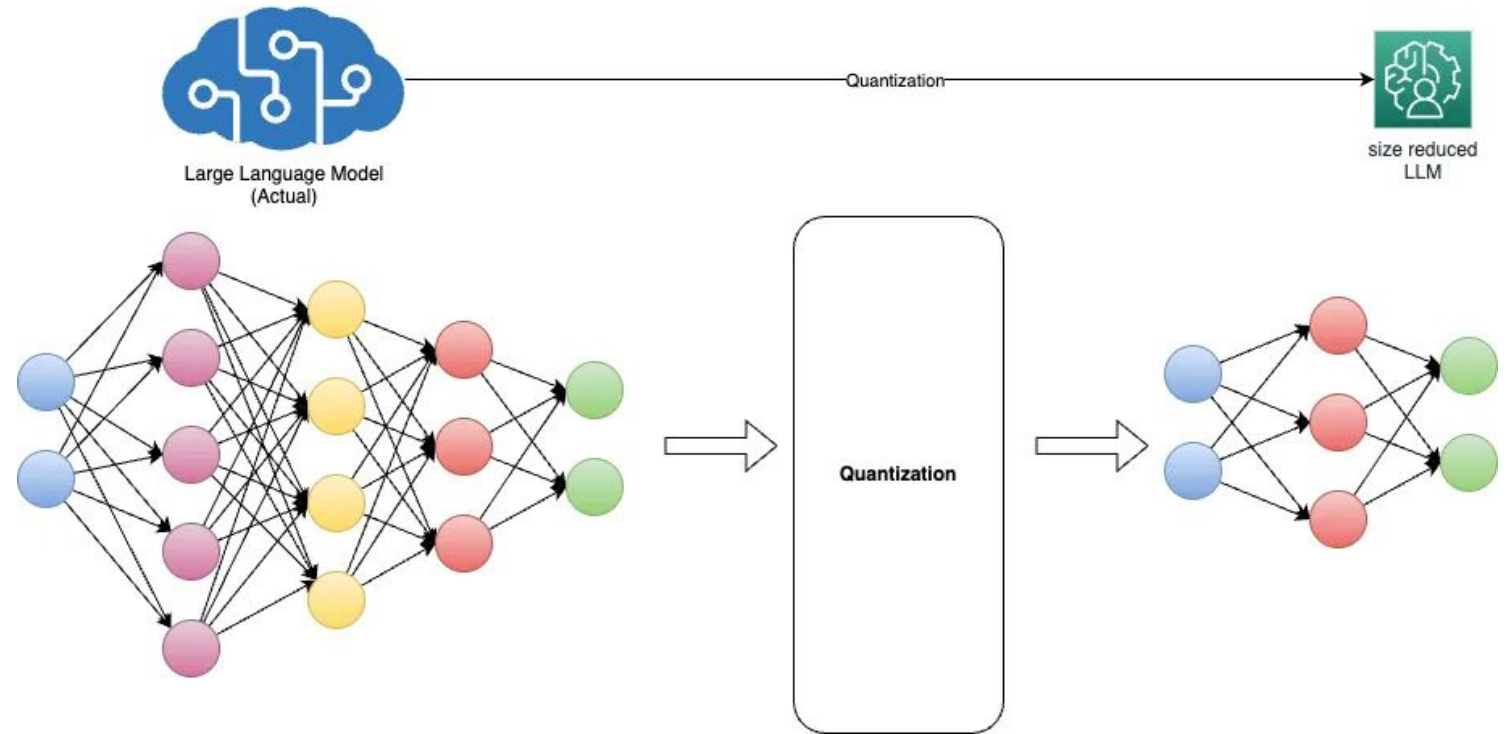
<https://www.gartner.com/en/articles/understand-and-exploit-gen-ai-with-gartner-s-new-impact-radar>

Source: Gartner  
© 2023 Gartner, Inc. and/or its affiliates. All rights reserved. 2683355



# Edge AI and Generative AI

- AI models optimizations
- Quantization
- Pruning
- Knowledge distillation

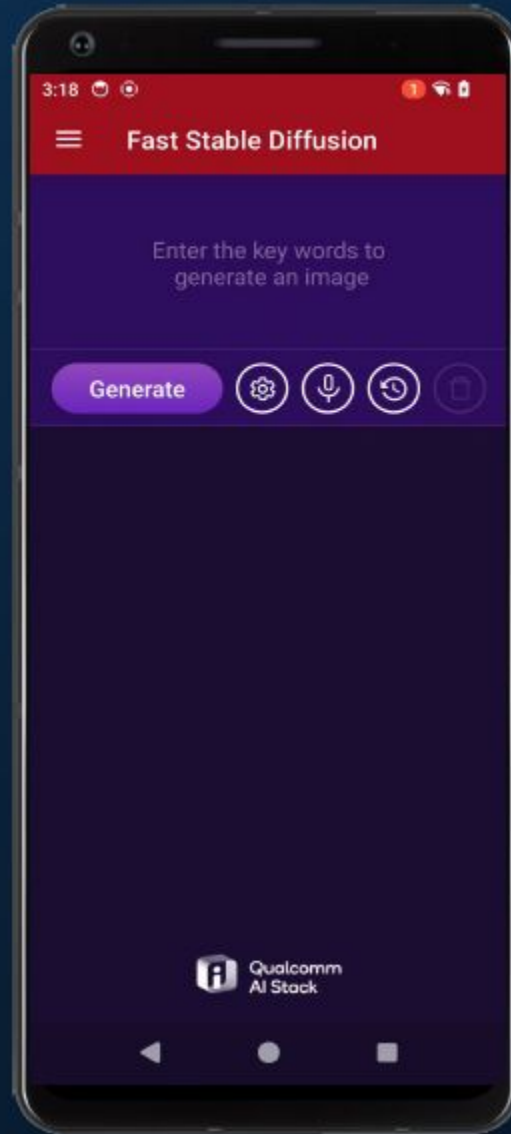


<https://int8.io/local-large-language-models-beginners-guide/>

<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

# Edge AI and Generative AI

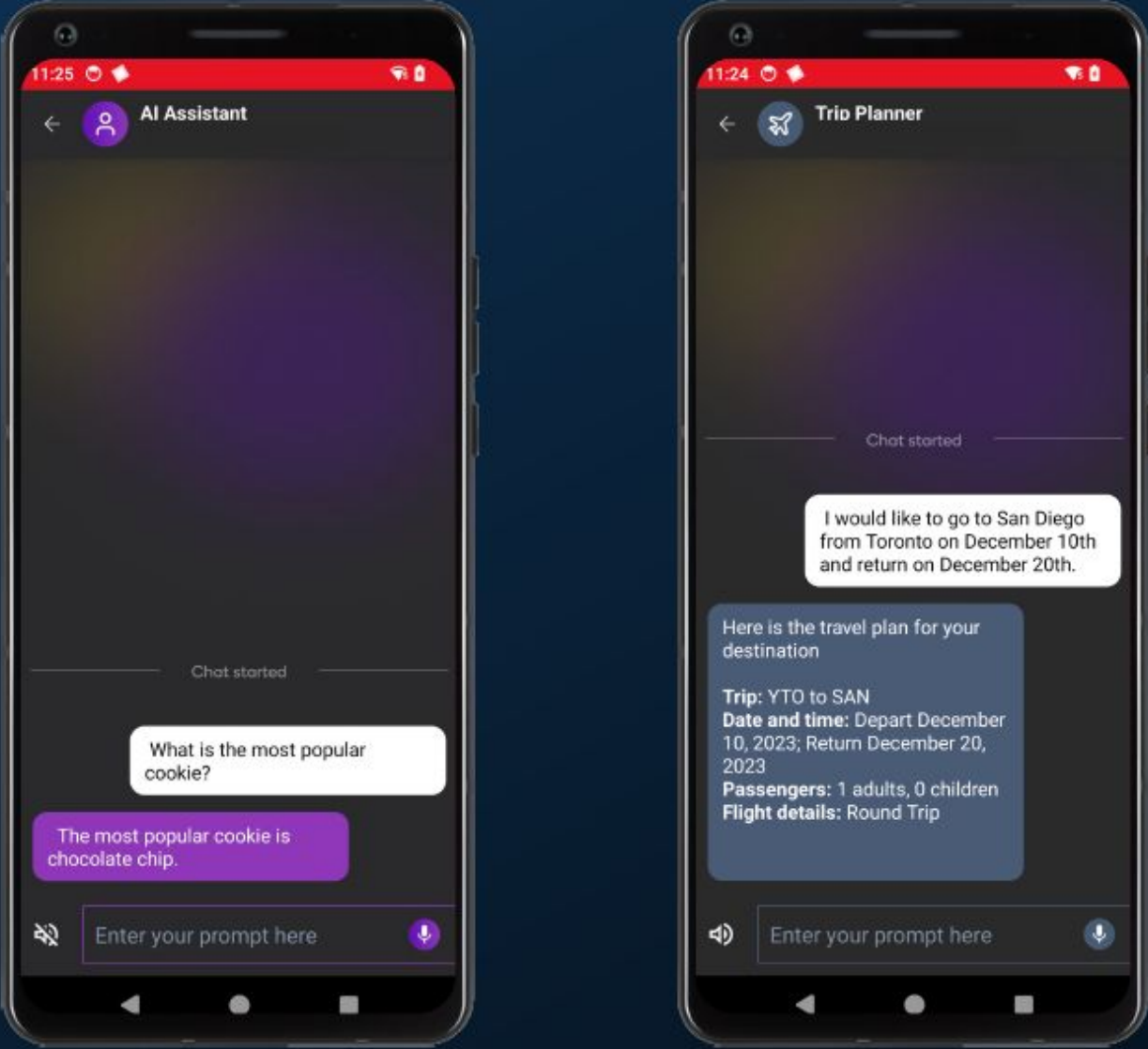
At Snapdragon Summit 2023

## World's fastest Llama 2-7B on a phone

Up to 20 tokens per second

Demonstrating both chat and application interaction on device

World's first demonstration of speculative decoding running on a phone



The image displays two smartphone screens side-by-side, demonstrating AI capabilities. The left screen shows the 'AI Assistant' interface with a chat conversation about cookies. The right screen shows the 'Trio Planner' interface with a chat conversation about a travel plan to San Diego.

**AI Assistant Chat:**

Chat started

What is the most popular cookie?

The most popular cookie is chocolate chip.

Enter your prompt here

**Trio Planner Chat:**

Chat started

I would like to go to San Diego from Toronto on December 10th and return on December 20th.

Here is the travel plan for your destination

Trip: YTO to SAN  
Date and time: Depart December 10, 2023; Return December 20, 2023  
Passengers: 1 adults, 0 children  
Flight details: Round Trip

Enter your prompt here

# Edge AI and Generative AI

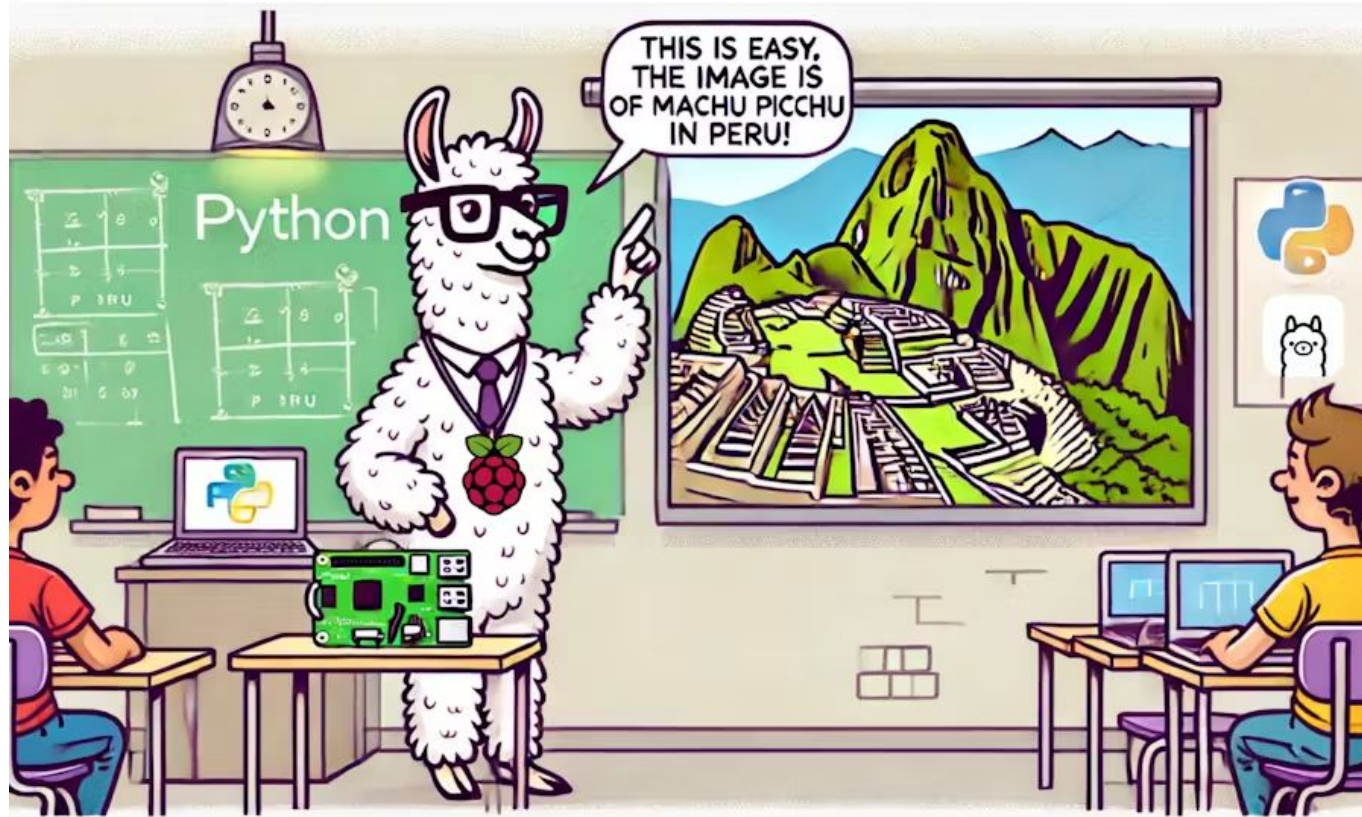
llama4micro 🐫🔬

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>



# Thanks

Prof. Jesús Alfonso López  
[jalopez@uao.edu.co](mailto:jalopez@uao.edu.co)

<https://www.linkedin.com/in/jesus-alfonso-lópez-sotelo-76100718/>

Universidad Autónoma de Occidente

