TECHNISCHE UNIVERSITÄT DRESDEN

DRESDEN concept

How Large Language Model train

Anna Poetsch

Research Group „Biomedical Genomics“, Biotechnology Center TU Dresden, NCT Dresden, and CSBD

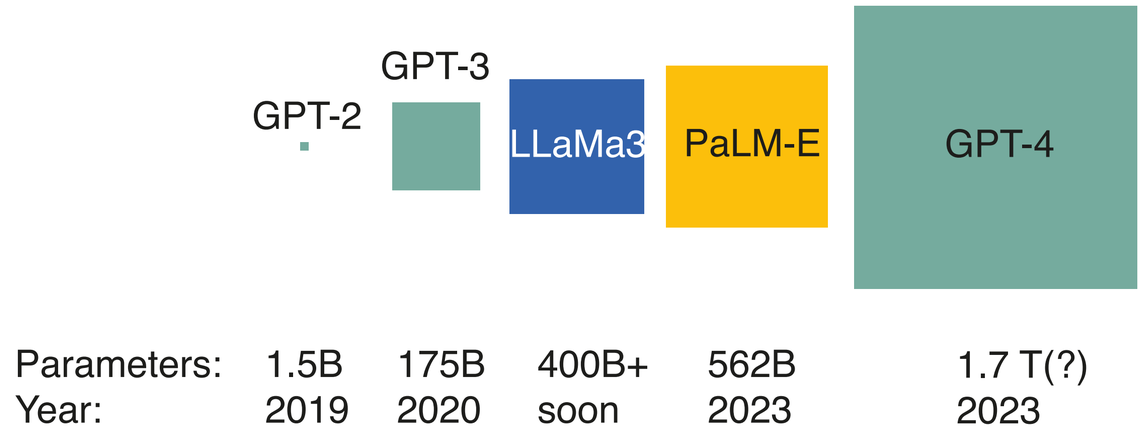
@apoetsch

Outline

- What are Large Language Models?
- Foundation model training
- Training a transformer model
- Fine-tuning Large Language Models
- Hallucination and bias
- Large Language Models with language-like data

Pre-training and fine-tuning

Foundation models train language on large corpora of data



BIG

expensive

rigid

Fine-tuning

- assistant
- image generation
- translation
- speech recognition
- ...

“DallE2, please give me an illustration of damaged DNA in comic style”:

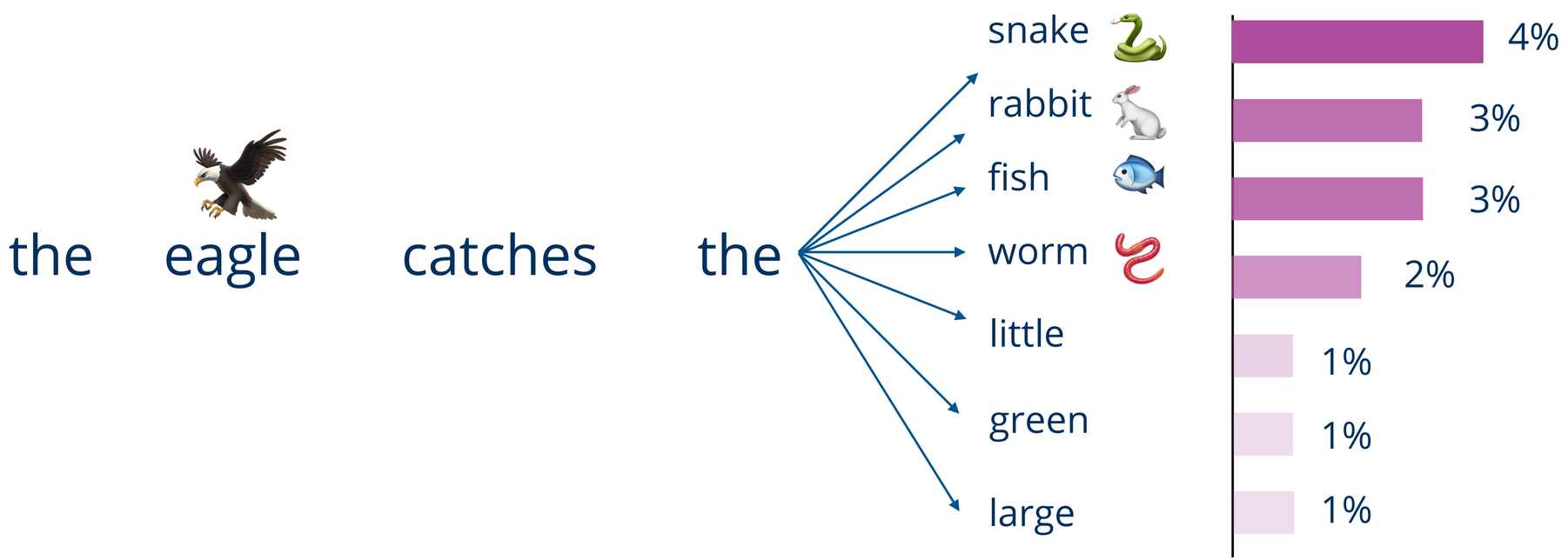
efficient

flexible

cheap

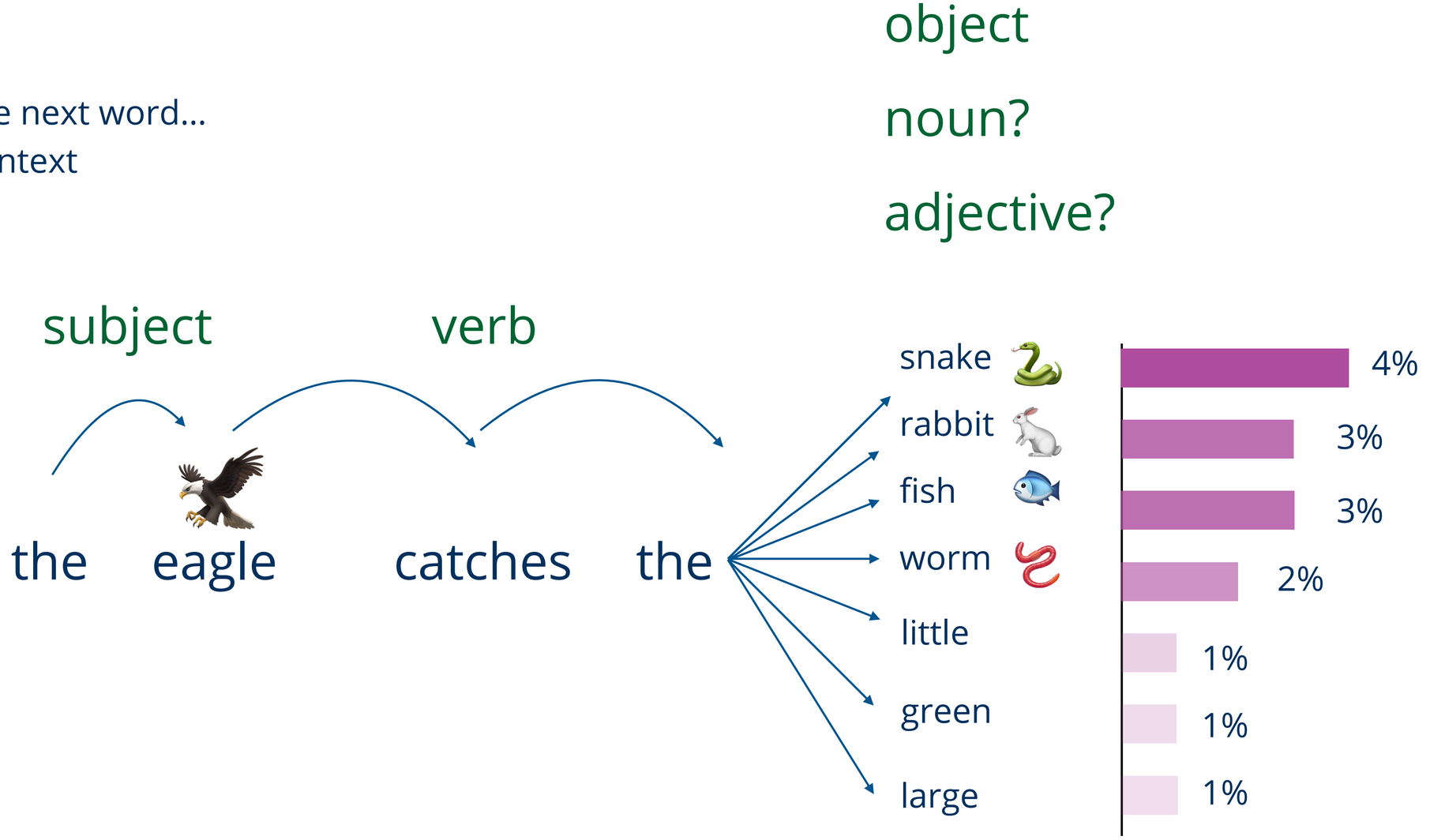
Large Language Models

Predicting the next word...



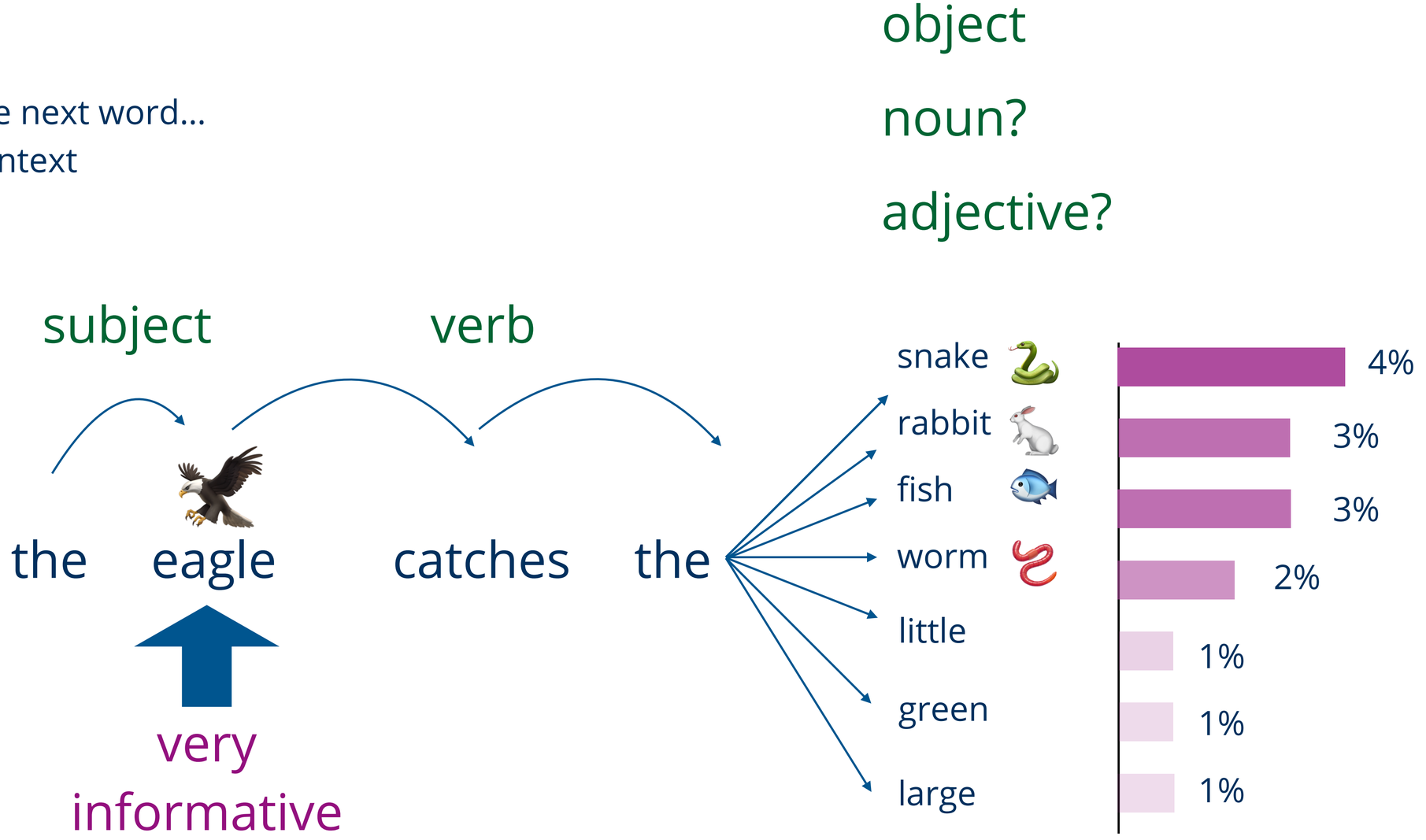
Large Language Models

Predicting the next word...
...requires context



Large Language Models

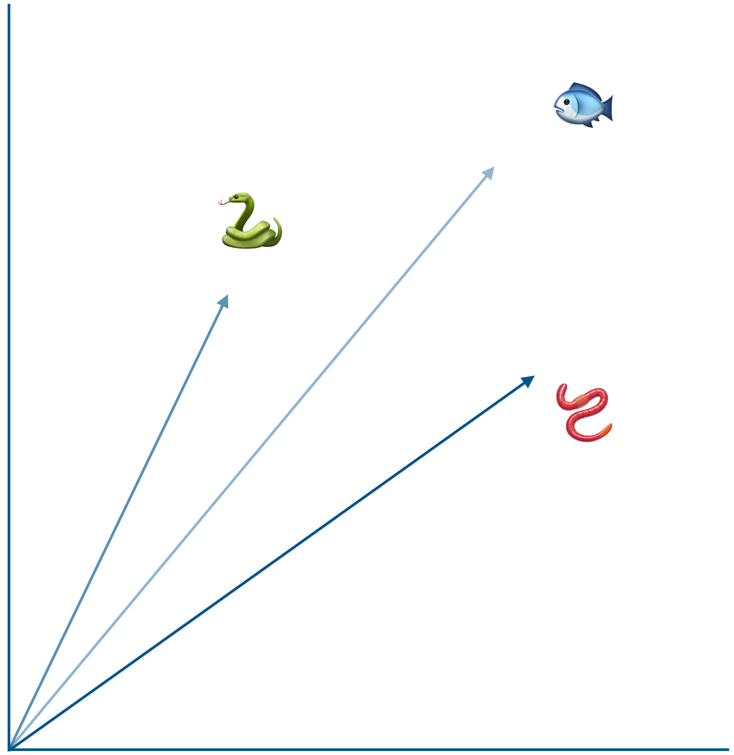
Predicting the next word...
...requires context



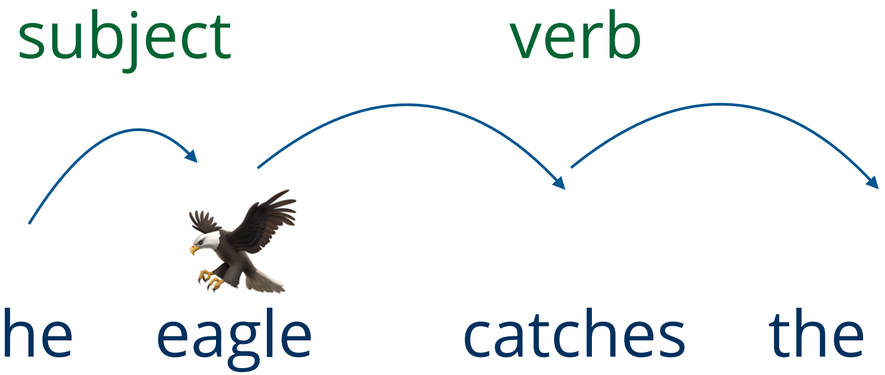
The embedding of words/ tokens

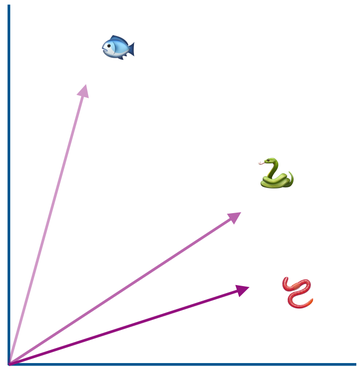
Tokens are assigned a vector, which places them into a multi-dimensional space

the	eagle	catches	the	???
5.4	3.2	0.2	5.4	
2.3	1.4	0.3	2.3	
1.2	0.4	5.6	1.2	
.	.	.	.	
.	.	.	.	
.	.	.	.	
6.4	1.2	3.2	6.4	
0.4	8.6	4.1	0.4	



Training a large language model





Input embedding

5.4	3.2	0.2	5.4
2.3	1.4	0.3	2.3
1.2	0.4	5.6	1.2
.	.	.	.
.	.	.	.
.	.	.	.
6.4	1.2	3.2	6.4
0.4	8.6	4.1	0.4

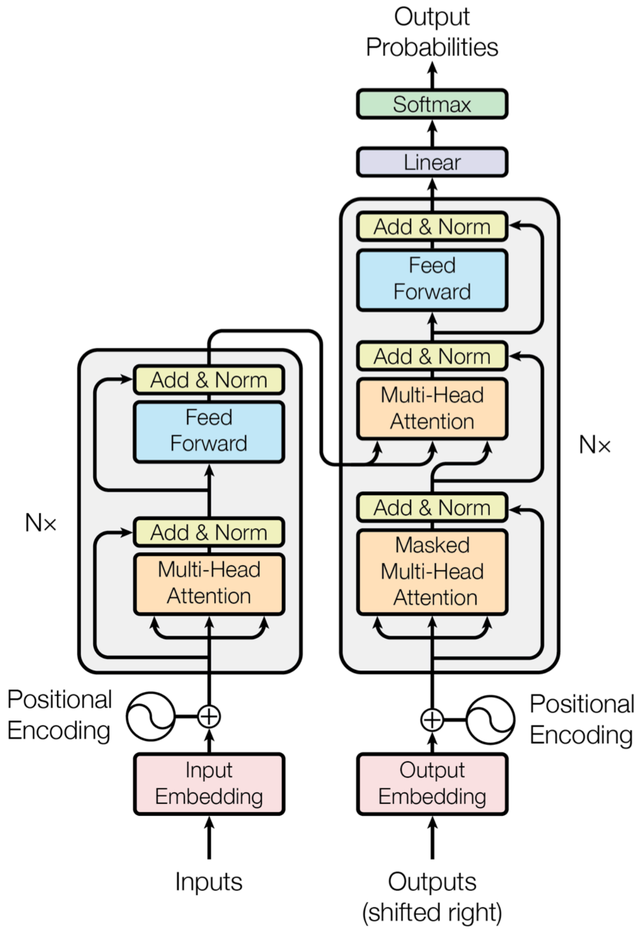
context

learned through attention blocks

Trained embedding

3.2	6.3	0.6	3.7
4.6	0.4	1.7	5.3
5.2	8.2	4.8	7.2
.	.	.	
.	.	.	
.	.	.	
5.9	3.2	3.9	2.1
0.3	0.4	6.2	0.9

“Attention is all you need”



GPT-3							Total parameters: 175,181,291,520 27,938 matrices
Embedding	12,288 d embed	*	50,257 n vocab				= 617,558,016
attention – Key	128 d query	*	12,288 d embed	* 96 n heads	* 96 n layers		= 14,495,514,624
attention – Query	128 d query	*	12,288 d embed	* 96 n heads	* 96 n layers		= 14,495,514,624
attention – Value	128 d query	*	12,288 d embed	* 96 n heads	* 96 n layers		= 14,495,514,624
Output	12,288 d embed	*	128 d value	* 96 n heads	* 96 n layers		= 14,495,514,624
Up-projection	49,152 n neurons	*	12,288 d embed	* 96 n layers			= 57,982,058,496
Down-projection	12,288 d embed	*	49,152 n neurons	* 96 n layers			= 57,982,058,496
Unembedding	50,257 n vocab	*	12,288 d embed				= 617,558,016

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

https://arxiv.org/abs/1706.03762

Hallucination, “making things up”

Anna Poetsch is a Computational Biologist of the TU Dresden.
She has received a grant from the European Research Council (ERC) and is one of the winners of the “Cluster of Excellence: Inflammation at Interfaces”. The Cluster of Excellence led by the TU Dresden and the University of Leipzig has been funded with more than 40 million euros by the German Federal Ministry of Education and Research (BMBF) for a period of ten years.

Bias

A model is only as good as the data used to train it.

Trained it with racism and sexism, it will return just that

Compensating for bias is hard.

Google Gemini: Adi Robertson / The Verge

Other text-like data

Proteins

DNA

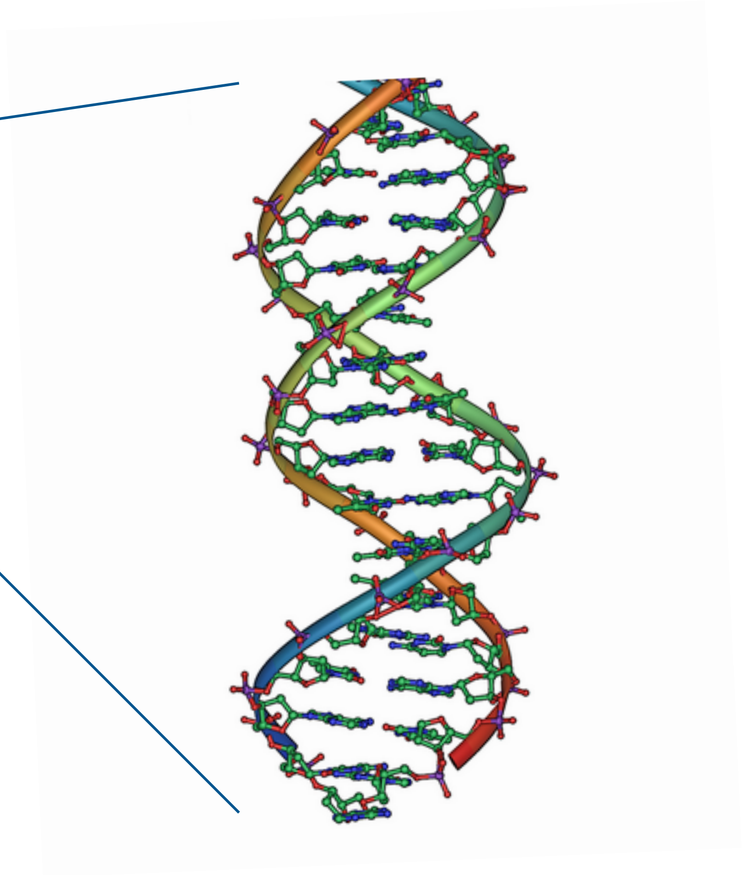
Sound

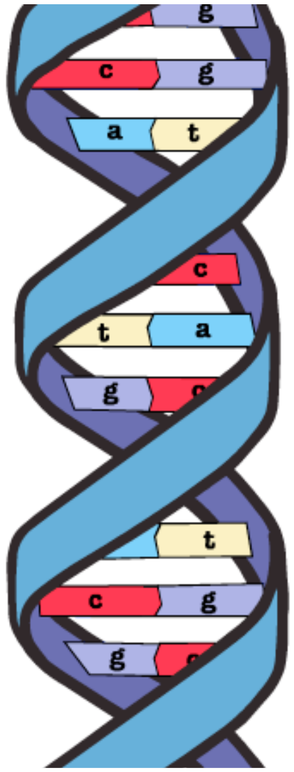
Music

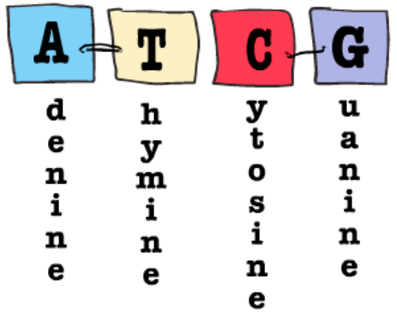
.

.

.







Summary, take-home messages, and further information

- Large Language models are models trained on next- or masked-token prediction

- They learn a sense of grammar and syntax, as well as language context

- They can be fine-tuned for a myriad of tasks (e.g. assistants and image generators)

- They are at risk of hallucination and amplifying bias

- Any data that resembles “language” can be used for training a model like this

Attention is all you need: https://arxiv.org/abs/1706.03762

How Transformers Work: A Detailed Exploration of Transformer Architecture: https://www.datacamp.com/tutorial/how-transformers-work

But what is a GPT? Visual intro to transformers: https://www.youtube.com/watch?v=wjZofJX0v4M

Attention in transformers, visually explained: https://www.youtube.com/watch?v=eMlx5fFNoYc