The Call of the **Transformer**

generated with Stable Diffusion **2.1**

NEURIPS map

https://neuripsav.vizhub.ai/

Why solving language?

“The limits of my language mean the limits of my world.” *

Ludwig Wittgenstein

* there’s a lot going on here as he was a somehow controversial fellow; assume we treat his words out of context

Shannon’s Guessing Game

Prediction and Entropy of Printed English, **1951**

R A T H E R – D R A M A T I C A L L Y – T H E – O T H E R – D A Y

Redundancy

Aoccdrnig to rscheearch at Cmabirgde Uinervtisy, it deosn't mttaer in waht oredr the ltteers in a wrod are, the olny iprmoetnt tihng is taht the frist and lsat ltteer be at the rghit pclae. The rset can be a ttoal mses and you can sitll raed it wouthit porbelm. Tihs is bcuseae the huamn mnid deos not raed ervey lteter by istlef, but the wrod as a wlohe. Amzanig huh?

Data-to-Information Ratio

Rules exist, but they get drowned by noise!

It’s all statistics

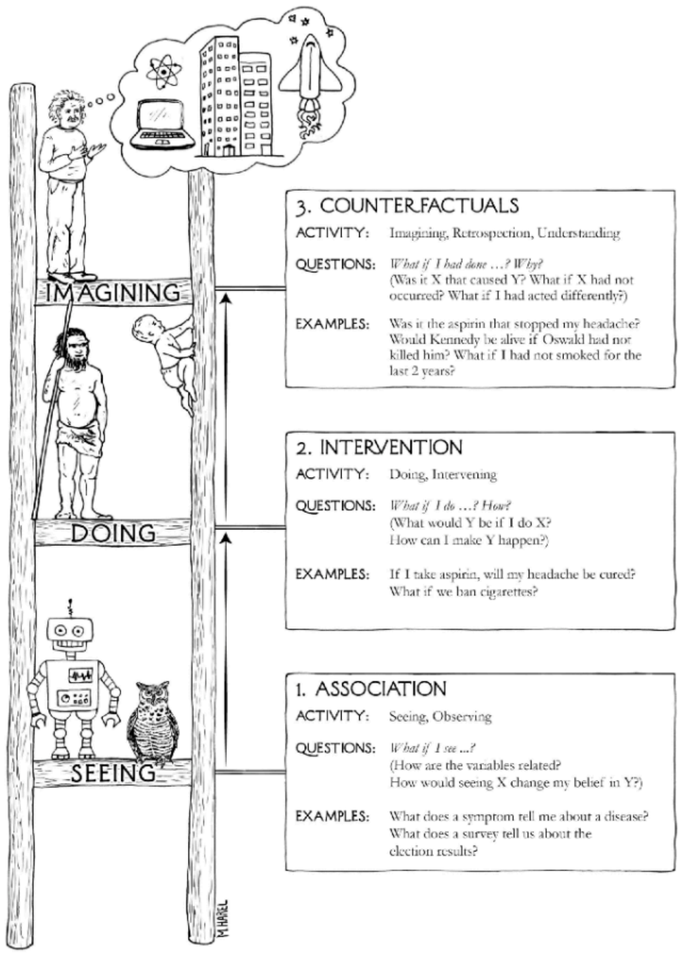
(but not **clever** statistics)

“With four parameters I can fit an elephant, and with five I can make him wiggle his trunk.”

John von Neumann

It’s all statistics

(but not **clever** statistics)



The Book of **Why**, Judea Pearl

The causes of data don’t lie in the data.

Interlude

“You cannot answer a question that you cannot ask, and you cannot ask a question that you have no words for.”

Judea Pearl

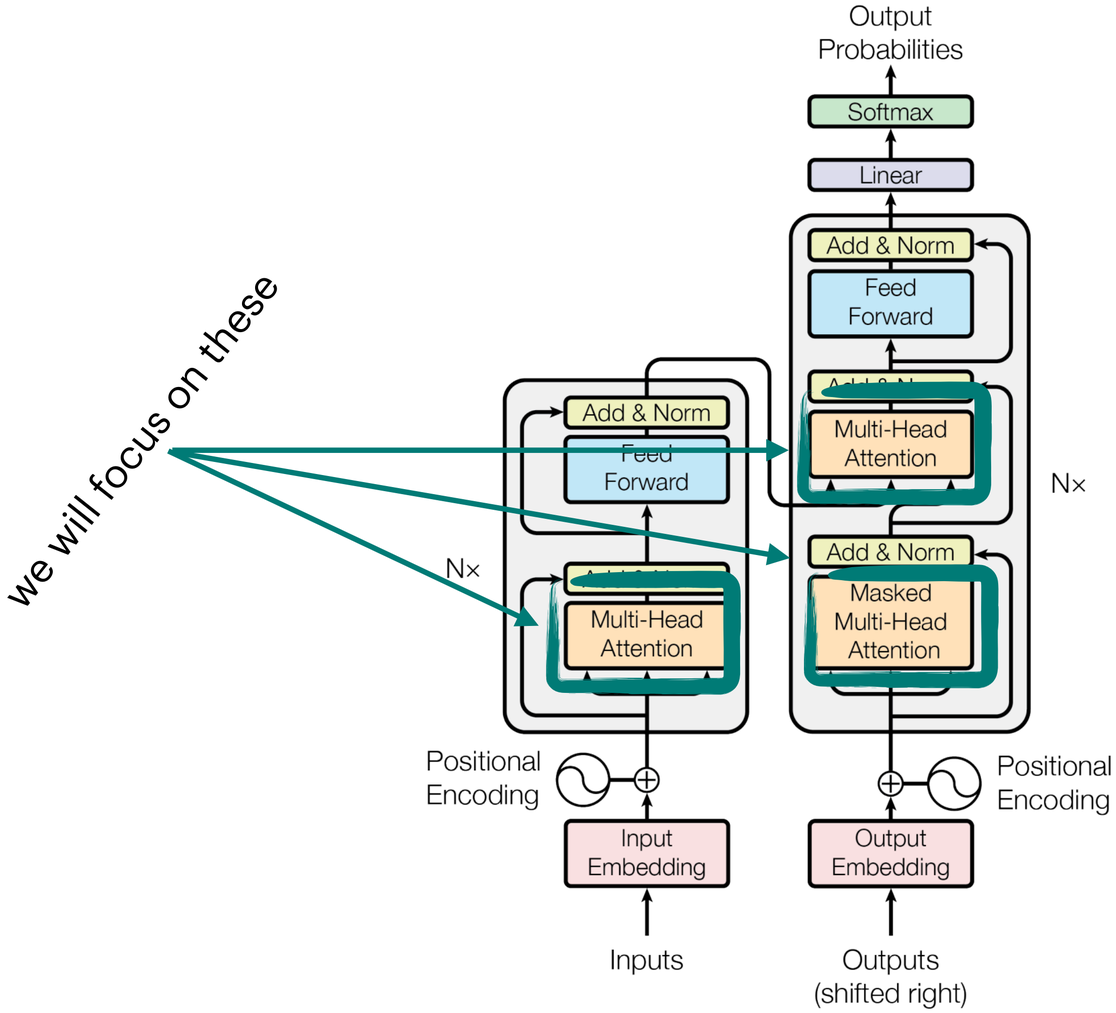
Transformer

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

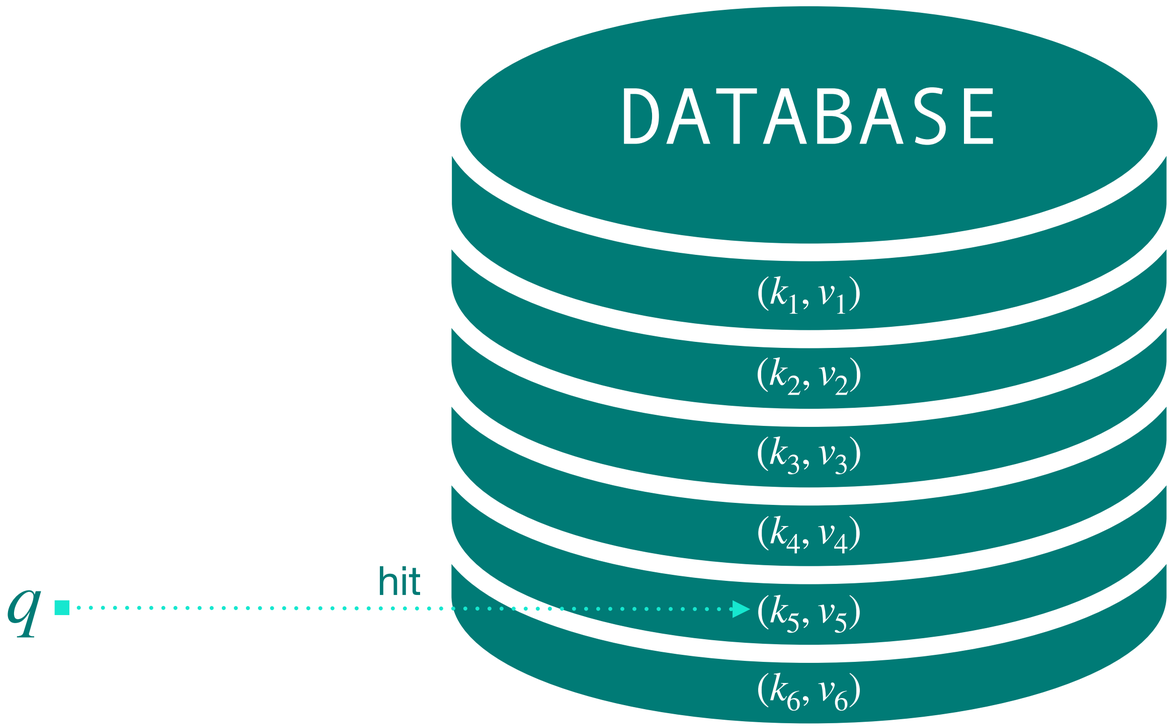
Abstract

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.



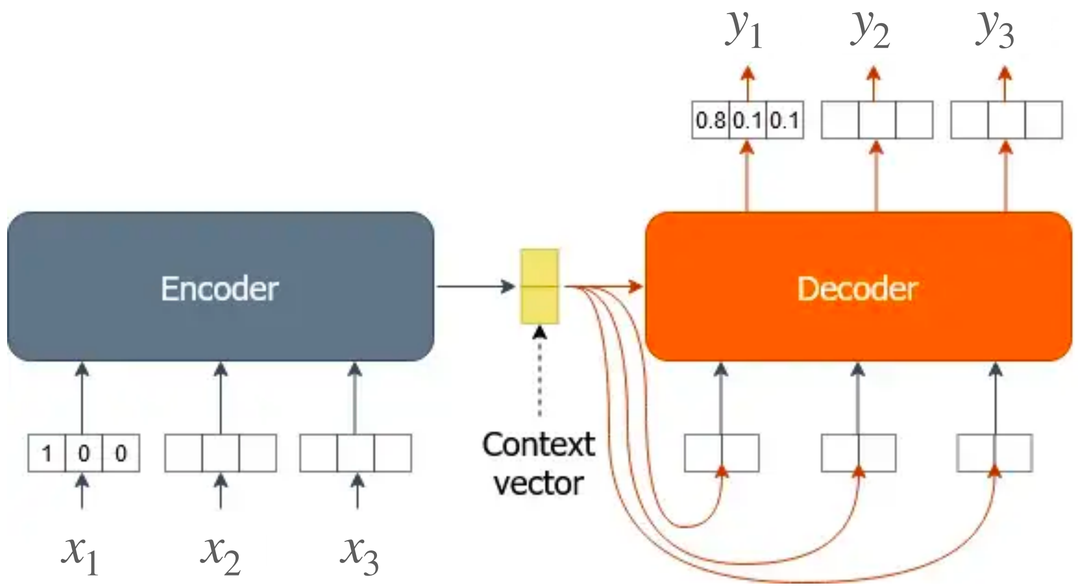
Attention

Why **q**ueries, **k**eys, and **v**alues?



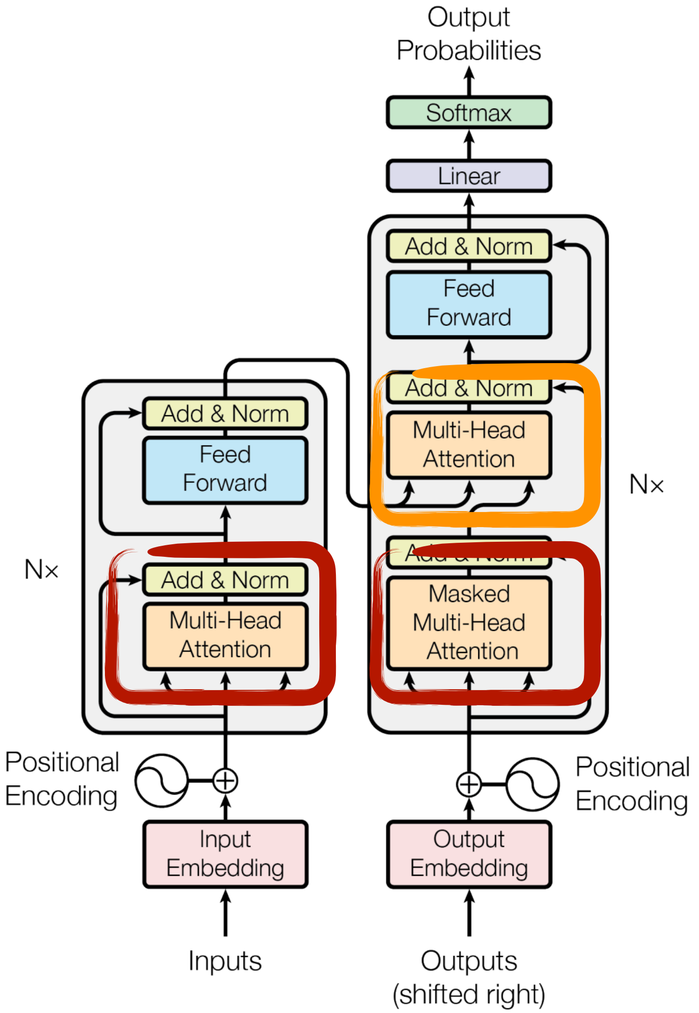
Attention

brief **seq2seq** (encoder-decoder) background



Attention

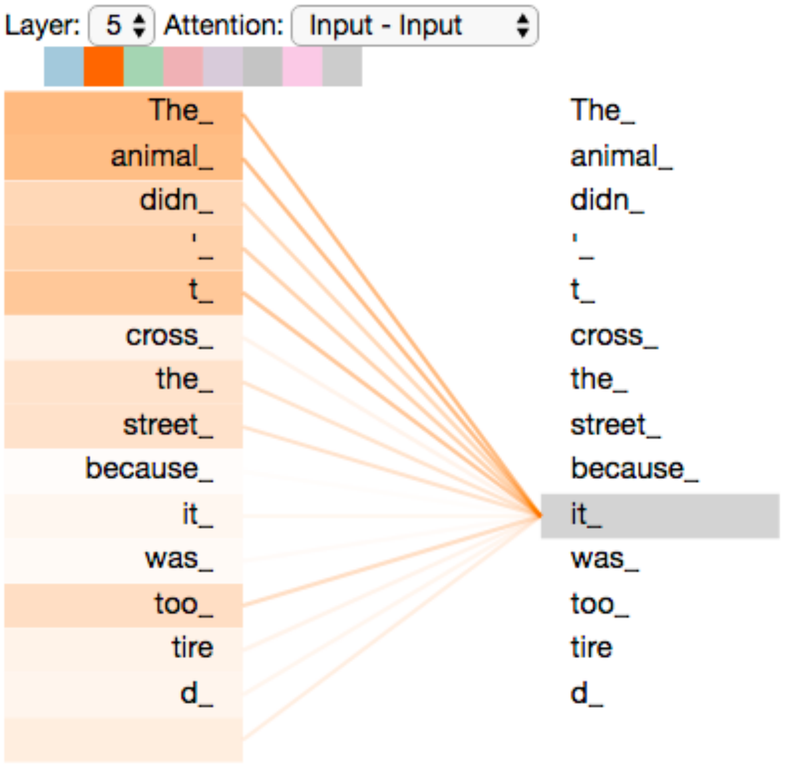
self-attention



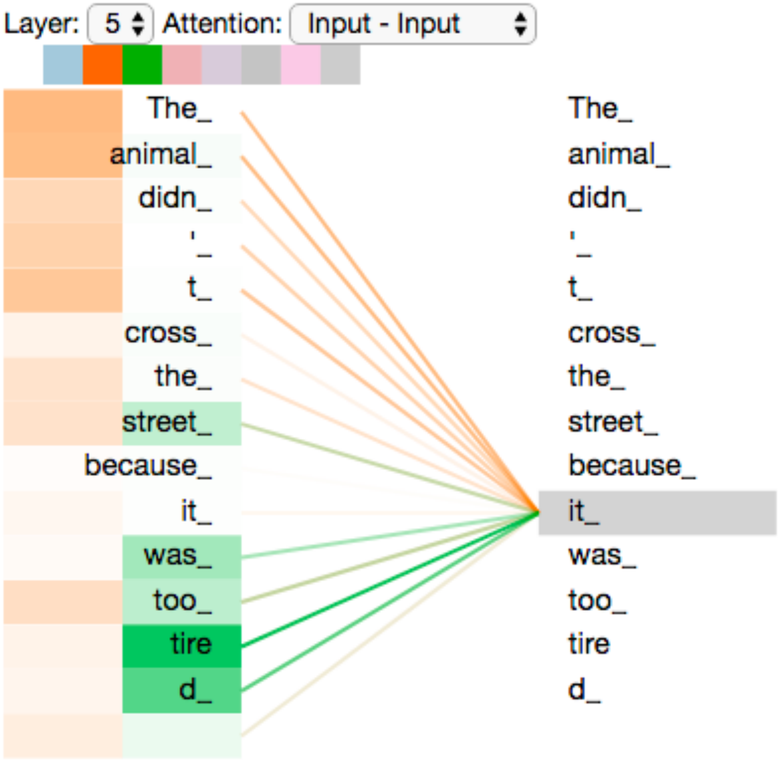
cross-attention

Attention

single-head

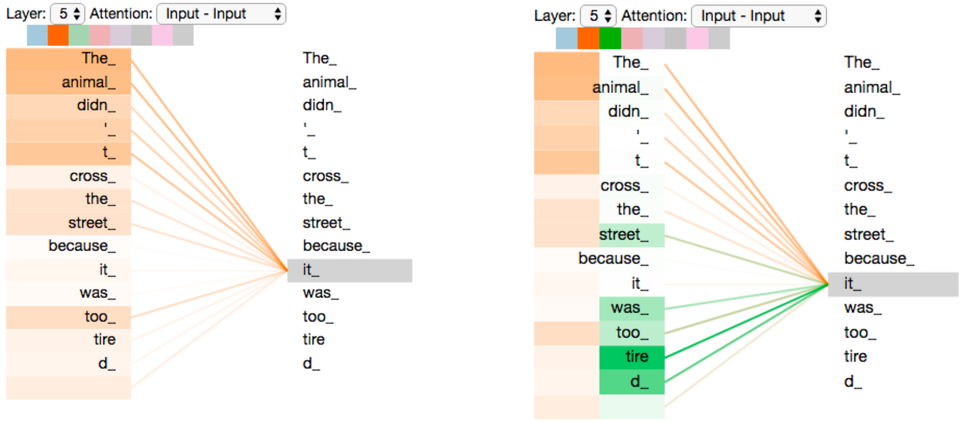


multi-head

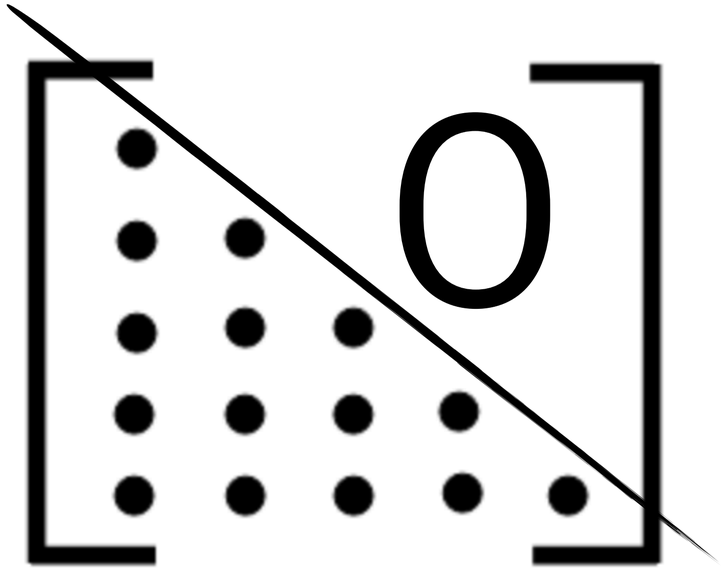


Attention

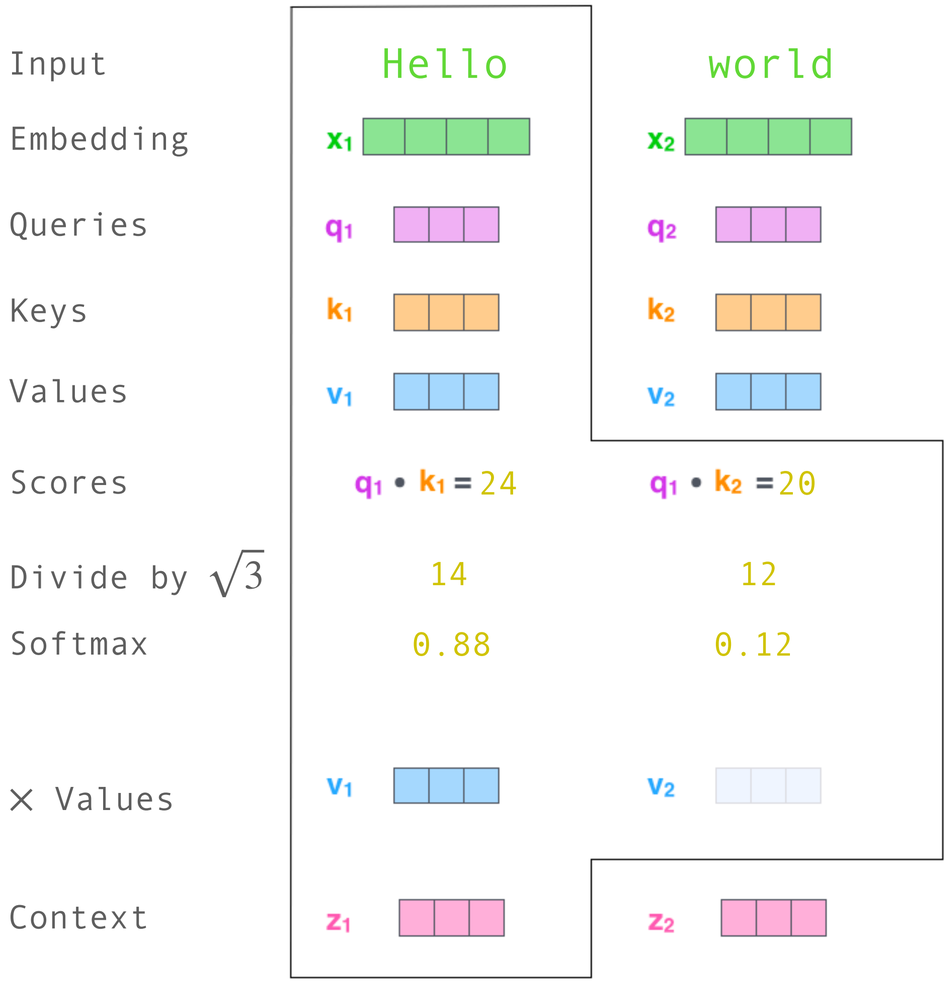
non-masked



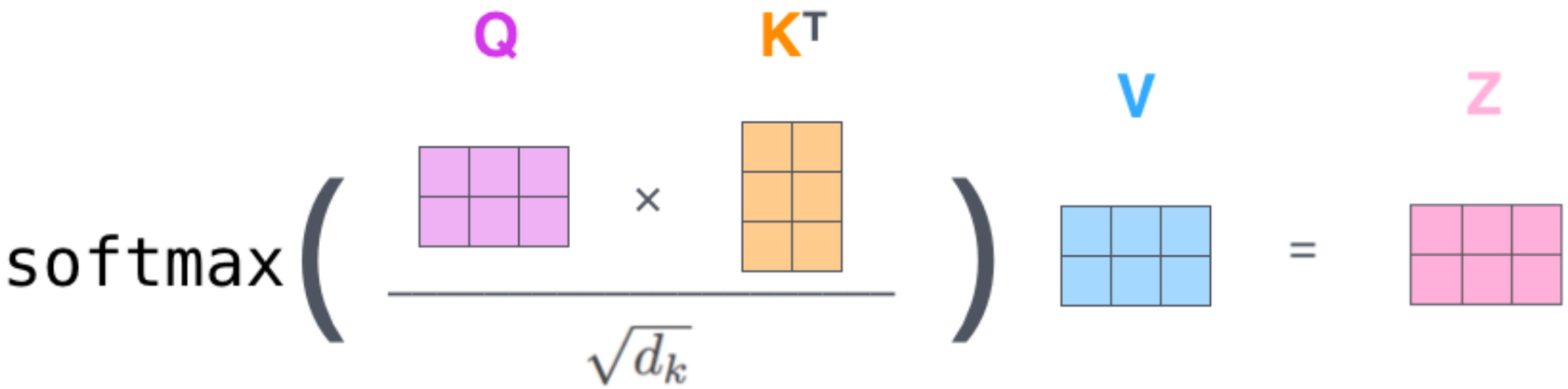
masked



Attention


Input
Hello
world
Embedding
x1
x2
Queries
q1
q2
Keys
k1
k2
Values
v1
v2
Scores
q1 • k1 = 24
q1 • k2 = 20
Divide by √3
14
12
Softmax
0.88
0.12
× Values
v1
v2
Context
z1
z2

Attention



Interlude

”The Transformer is a magnificent neural network architecture because it is a general-purpose differentiable computer.”

Andrej Karpathy

Status Quo

huggingface/transformers

174 models

Language Models

Auto**regressive**

THE CAT SITS ON THE[…]

GPT

Auto**encoding**

THE […] SITS ON THE SOFA.

BERT

Scaling Laws

DeepMind

Training Compute-Optimal Large Language Models

Jordan Hoffmann*, Sebastian Borgeaud*, Arthur Mensch*, Elena Buchatskaya, Trevor Cai, Eliza Rutherford,
Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland,
Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan,
Erich Elsen, Jack W. Rae, Oriol Vinyals and Laurent Sifre*
*Equal contributions

Given a specific compute budget, **how big** of a model should we train and for **how many tokens**?

The Players

Chinchilla	PaLM	MT-NLG	Gopher	GPT3	LaMDA
2022	2022	2021	2021	2020	2022
70B	540B	530B	280B	175B	137B

The Law

$$L(N, D) = \frac{A}{N^{\alpha}} + \frac{B}{D^{\beta}} + E$$

finite model

finite dataset

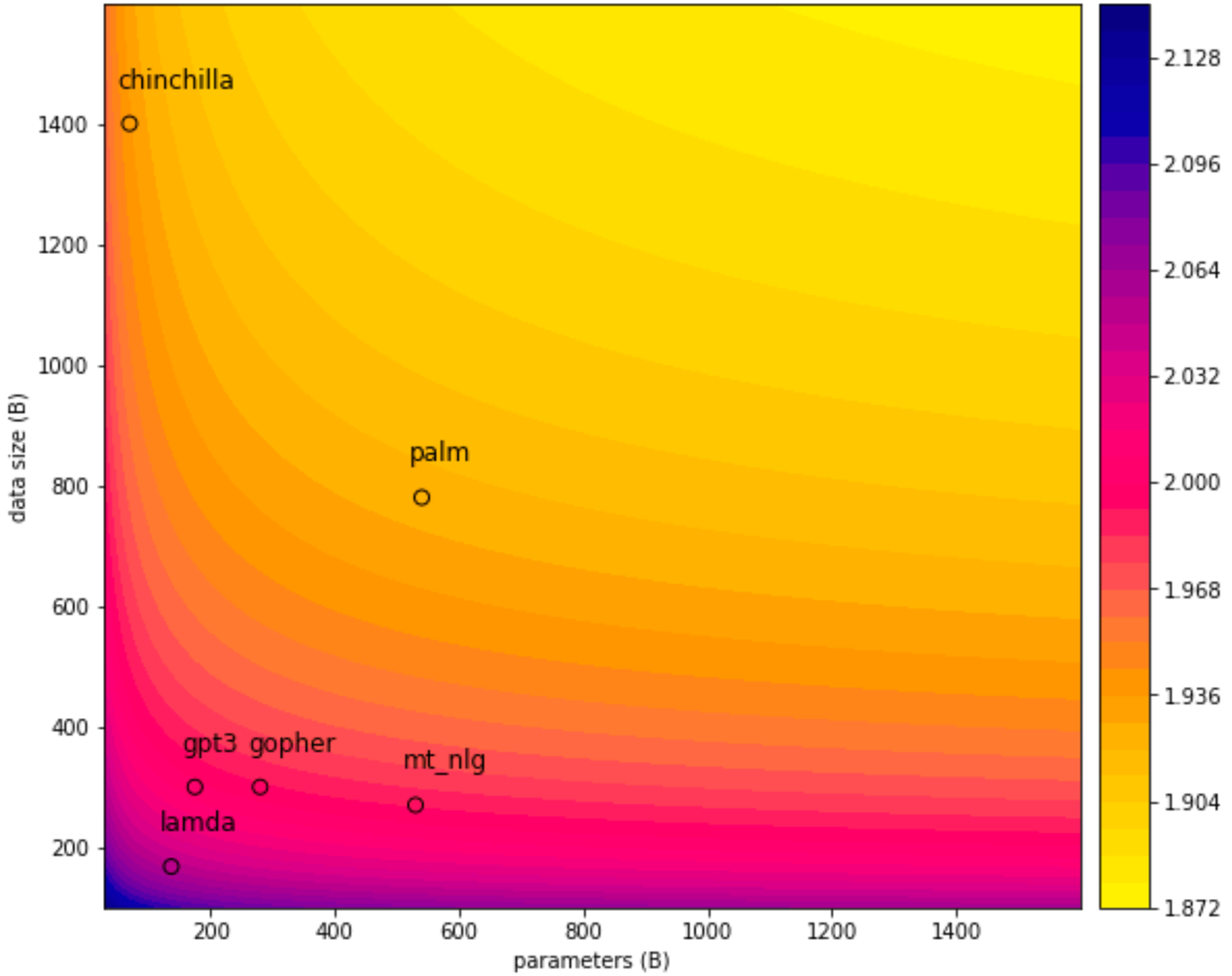
irreducible

The Law

Constants for the **MassiveText** dataset: **10.5 TB** of text

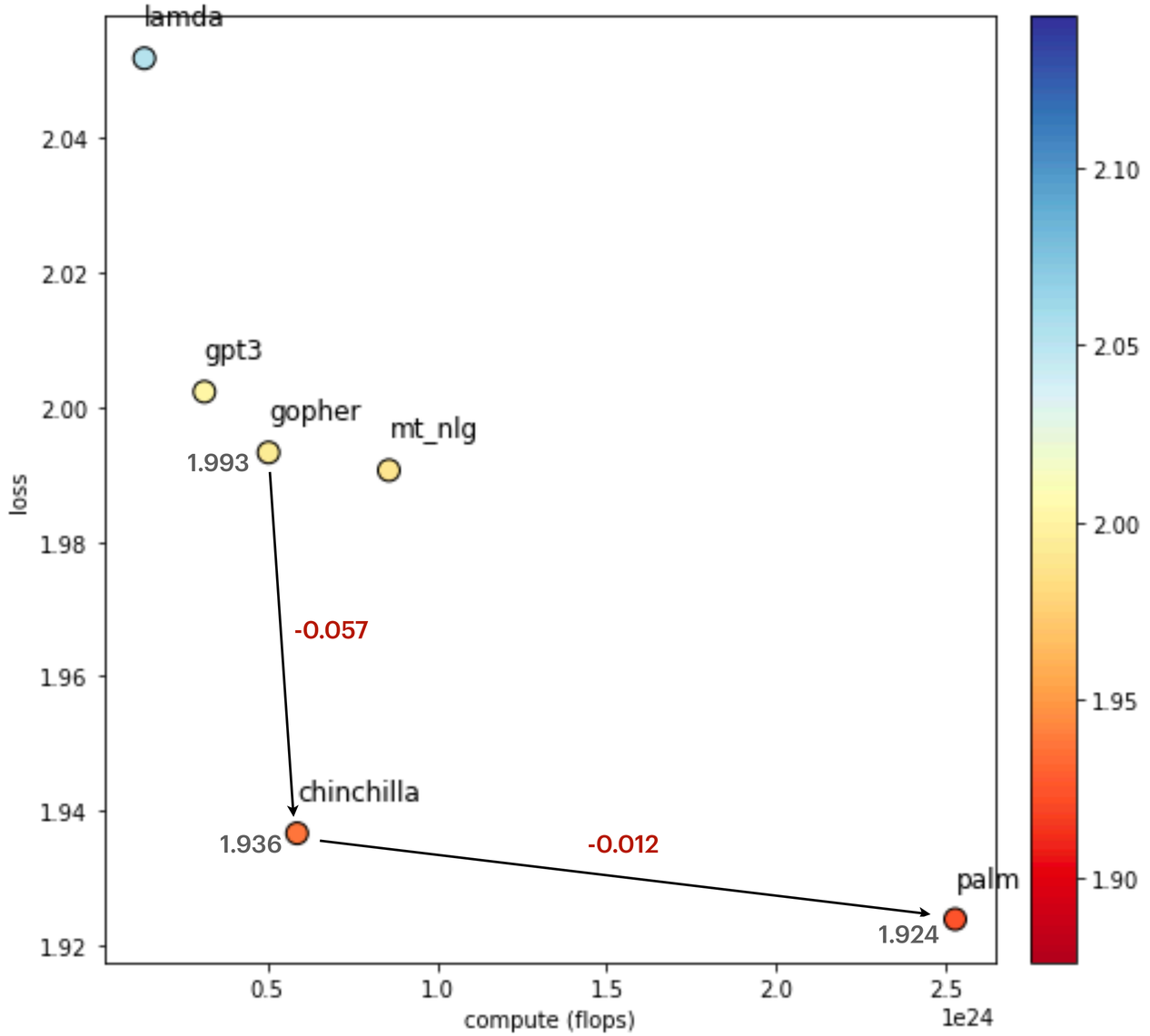
$$L(N, D) = \frac{406.4}{N^{0.34}} + \frac{410.7}{D^{0.28}} + 1.69$$

The Law



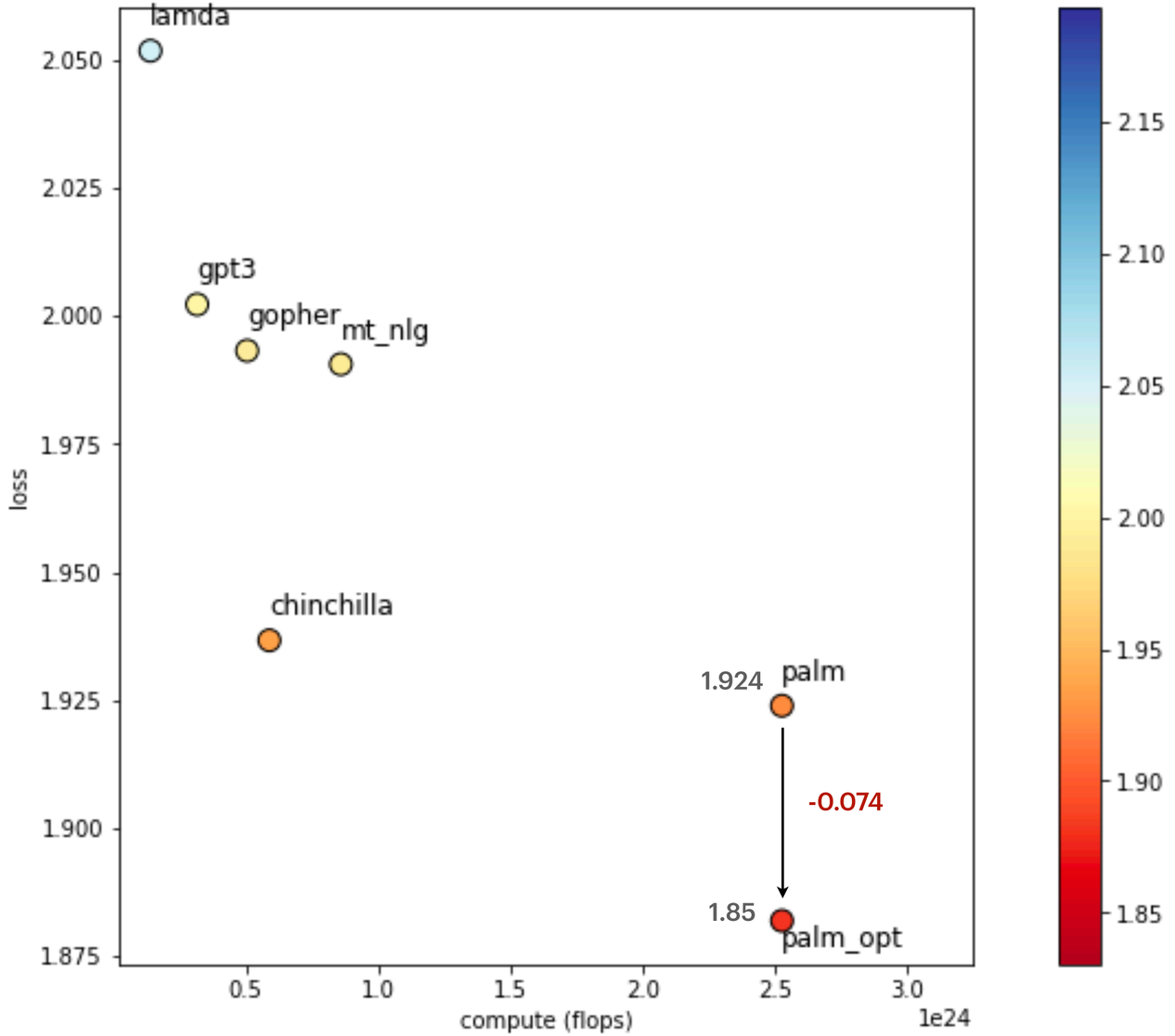
nostalgebraist, chinchilla's wild implications

The Law



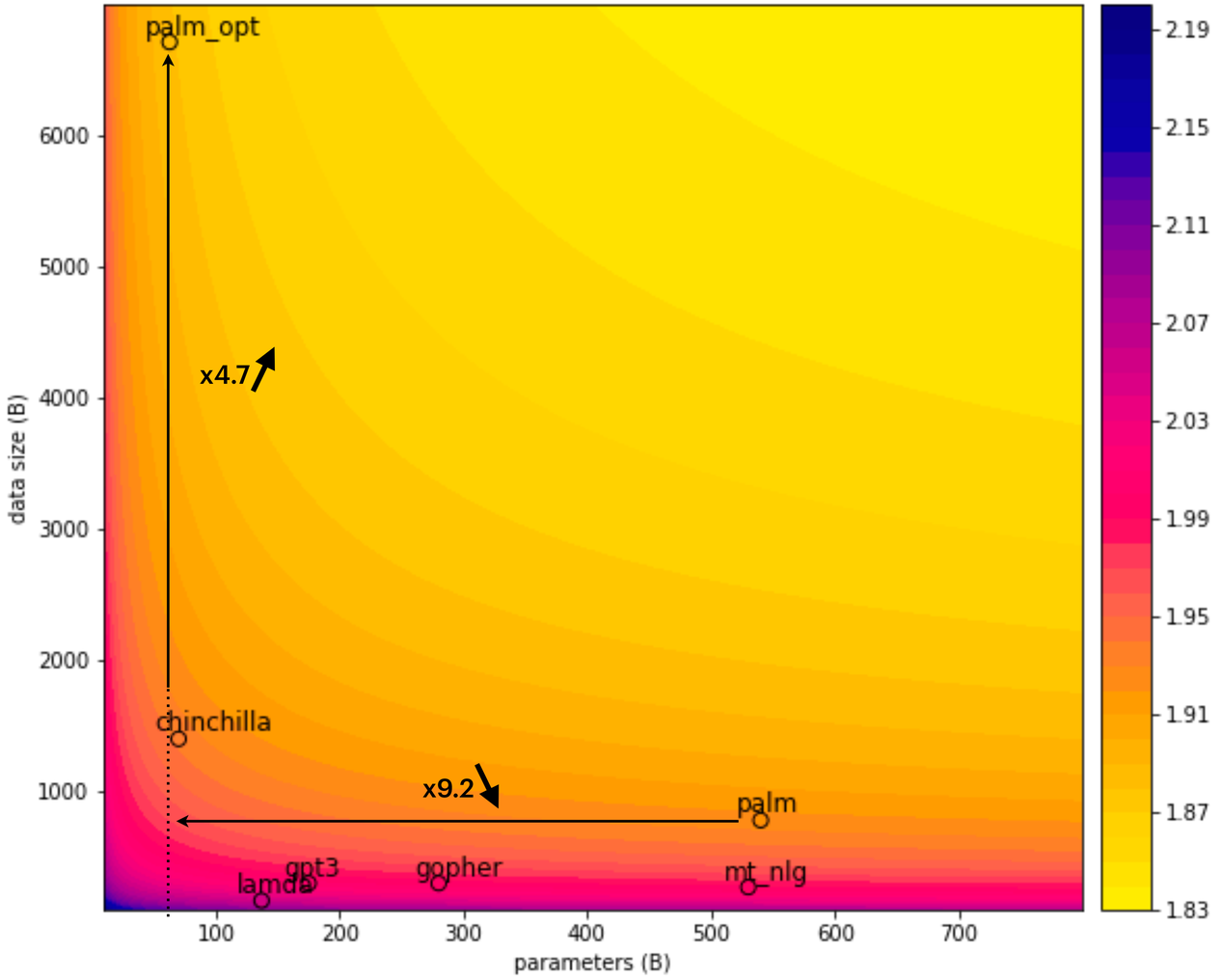
nostalgebraist, chinchilla's wild implications

The Law



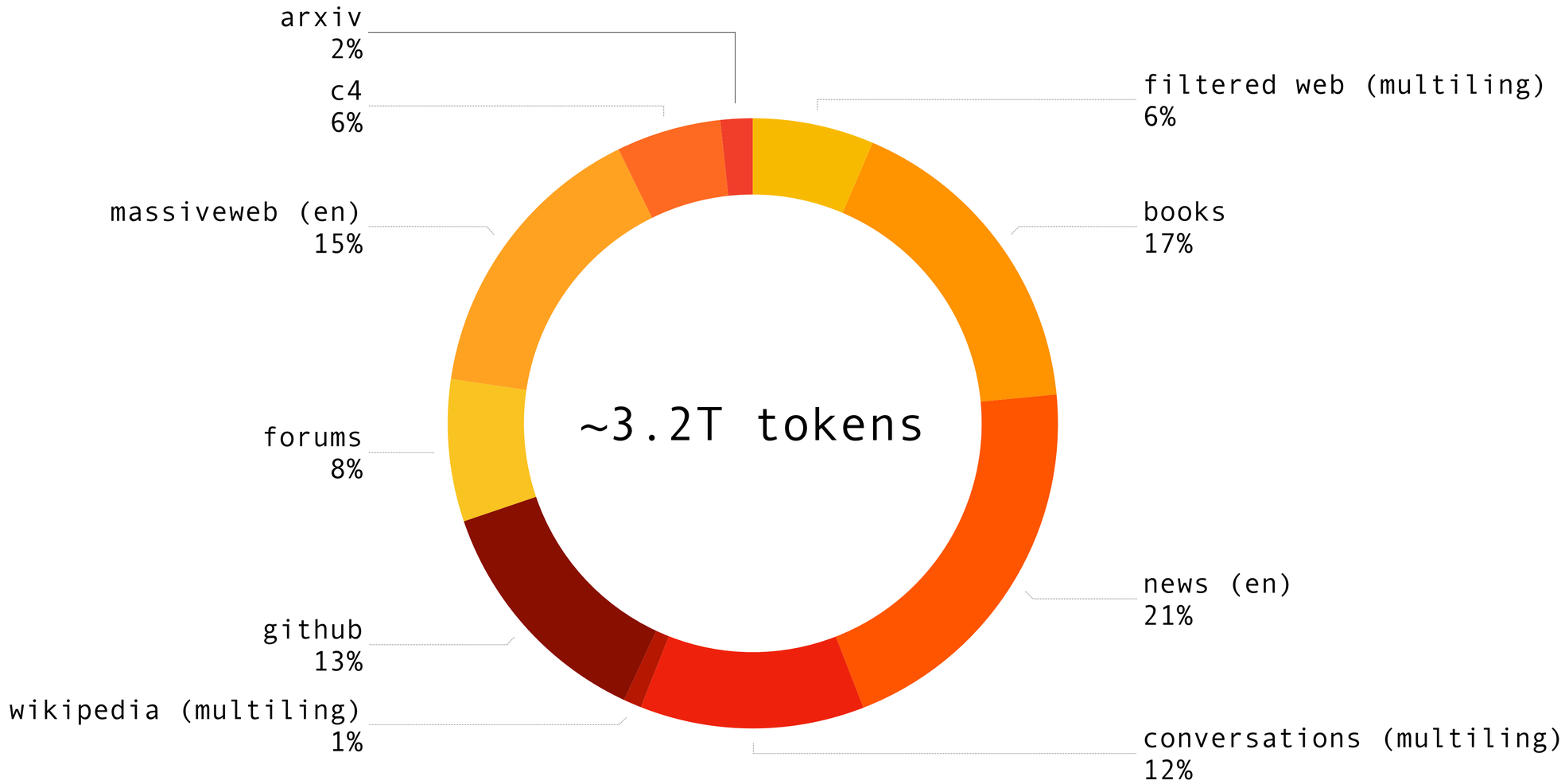
nostalgebraist, chinchilla's wild implications

The Law



nostalgebraist, chinchilla's wild implications

Data



nostalgebraist, chinchilla's wild implications

Future

relating multiple texts to each other

numbers and math

a notion of time

rare events

data hunger

Thanks for your