CSCI E-11

Class 10

April 7, 2026

Generative AI, Attention, and Transformers

Harvard
Extension School
HARVARD DIVISION OF
CONTINUING EDUCATION

News

Is encryption a solved problem (other than quantum)?

https://teleguard.com

404 Media’s investigation of TeleGuard:

A Secure Chat App’s Encryption Is So Bad It Is ‘Meaningless’

TeleGuard is an app downloaded more a million times that markets itself as a secure way to chat. The app uploads users’ private keys to the company’s server, and makes decryption of messages trivial.

JOSEPH COX · APR 2, 2026 AT 9:47 AM

https://www.404media.co/a-secure-chat-apps-encryption-is-so-bad-it-is-meaningless/

News story of the week

How A.I. Helped One Man (and His Brother) Build a $1.8 Billion Company

Who needs more than two employees when artificial intelligence can do so many corporate tasks? It's super efficient — and a little bit lonely.

New York Times,
April 2, 2026

The rest of the story…

Aakash Gupta
@aakashgupta

Subscribe

The NYT just profiled Medvi as the AI success story of the decade. $1.8B in projected revenue, 2 employees, Sam Altman's prediction made real.

Here's what the NYT didn't lead with.

Medvi received FDA Warning Letter #721455 in February 2026 for misbranding violations. Their clinician network OpenLoop suffered a data breach in January 2026 that exposed 1.6 million patient records. Futurism reported they used AI-generated deepfake before-and-after photos in their marketing. A class action lawsuit was filed in Delaware in November 2025. And now they're running 800+ fake doctor accounts on Facebook to sell compounded GLP-1s.

Last Week’s Reading

Last Week's Reading

ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

Ilya Sutskever
University of Toronto
ilya@cs.utoronto.ca

Geoffrey E. Hinton
University of Toronto
hinton@cs.utoronto.ca

Abstract

We trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 37.5% and 17.0% which is considerably better than the previous state-of-the-art. The neural network, which has 60 million parameters and 650,000 neurons, consists of five convolutional layers, some of which are followed by max-pooling layers, and three fully-connected layers with a final 1000-way softmax. To make training faster, we used non-saturating neurons and a very efficient GPU implementation of the convolution operation. To reduce overfitting in the fully-connected layers we employed a recently-developed regularization method called "dropout" that proved to be very effective. We also entered a variant of this model in the ILSVRC-2012 competition and achieved a winning top-5 test error rate of 15.3%, compared to 26.2% achieved by the second-best entry.

1 Introduction

Current approaches to object recognition make essential use of machine learning methods. To improve their performance, we can collect larger datasets, learn more powerful models, and use better techniques for preventing overfitting. Until recently, datasets of labeled images were relatively small — on the order of tens of thousands of images (e.g., NORB [16], Caltech-101/256 [8, 9], and CIFAR-10/100 [12]). Simple recognition tasks can be solved quite well with datasets of this size, especially if they are augmented with label-preserving transformations. For example, the current-best error rate on the MNIST digit-recognition task (<0.3%) approaches human performance [4]. But objects in realistic settings exhibit considerable variability, so to learn to recognize them it is necessary to use much larger training sets. And indeed, the shortcomings of small image datasets have been widely recognized (e.g., Pinto et al. [21]), but it has only recently become possible to collect labeled datasets with millions of images. The new larger datasets include LabelMe [23], which consists of hundreds of thousands of fully-segmented images, and ImageNet [6], which consists of over 15 million labeled high-resolution images in over 22,000 categories.

To learn about thousands of objects from millions of images, we need a model with a large learning capacity. However, the immense complexity of the object recognition task means that this problem cannot be specified even by a dataset as large as ImageNet, so our model should also have lots of prior knowledge to compensate for all the data we don't have. Convolutional neural networks (CNNs) constitute one such class of models [16, 11, 13, 18, 15, 22, 26]. Their capacity can be controlled by varying their depth and breadth, and they also make strong and mostly correct assumptions about the nature of images (namely, stationarity of statistics and locality of pixel dependencies). Thus, compared to standard feedforward neural networks with similarly-sized layers, CNNs have much fewer connections and parameters and so they are easier to train, while their theoretically-best performance is likely to be only slightly worse.

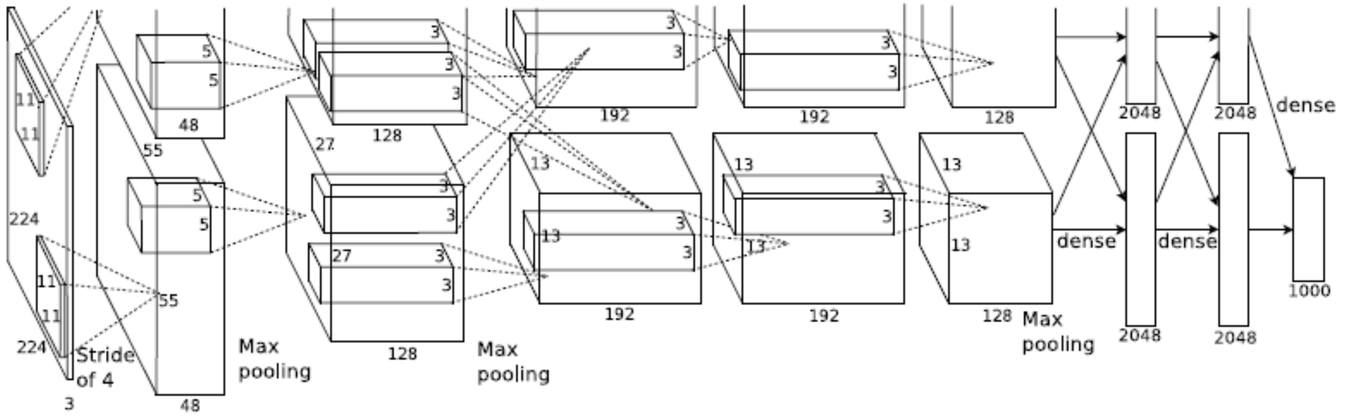
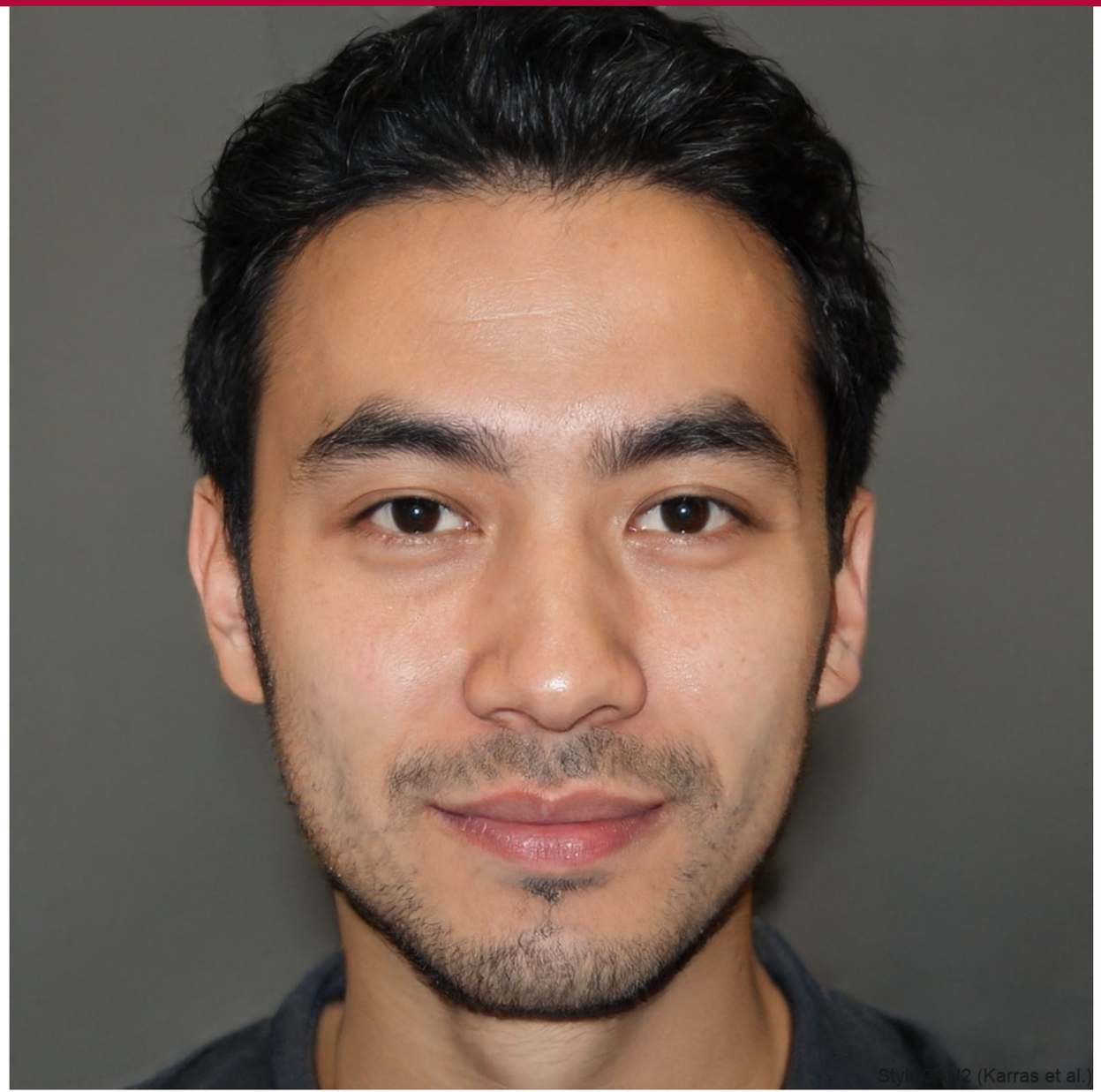


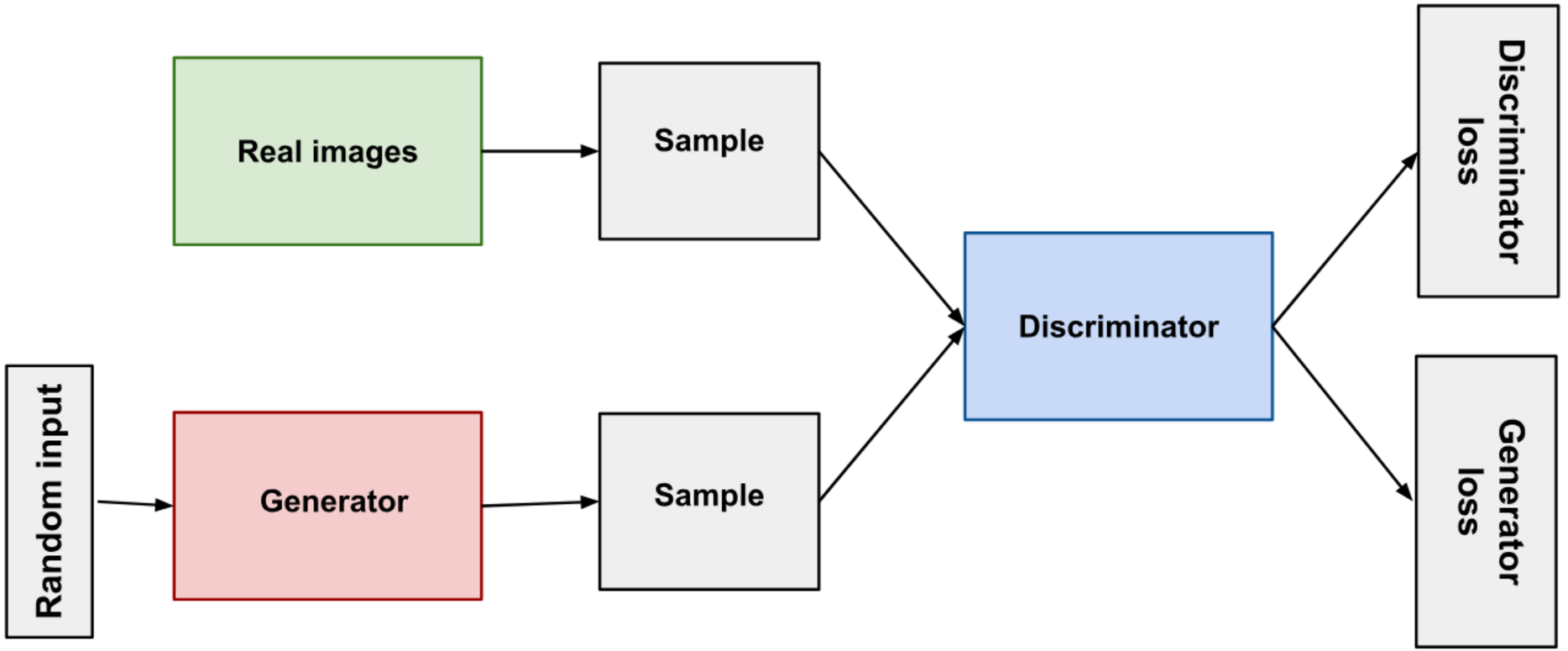
Figure 2: An illustration of the architecture of our CNN, explicitly showing the delineation of responsibilities between the two GPUs. One GPU runs the layer-parts at the top of the figure while the other runs the layer-parts at the bottom. The GPUs communicate only at certain layers. The network's input is 150,528-dimensional, and the number of neurons in the network's remaining layers is given by 253,440–186,624–64,896–64,896–43,264–4096–4096–1000.

Generative AI

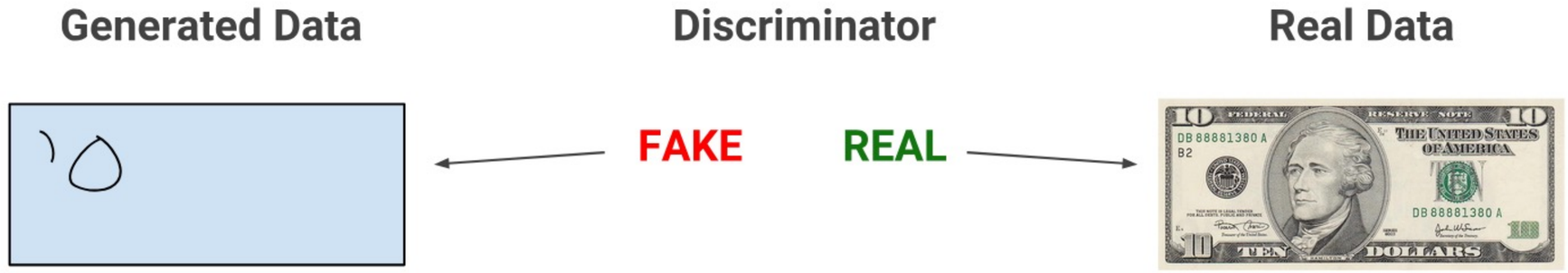
This Person Does Not Exist.



GANs


Real images
Sample
Random input
Generator
Sample
Discriminator
Discriminator loss
Generator loss

The generator and discriminator are in a race to beat each other.



As training progresses, the generator gets closer to producing output that can fool the discriminator:

Finally, if generator training goes well, the discriminator gets worse at telling the difference between real and fake. It starts to classify fake data as real, and its accuracy decreases.

Progress in GANs in late 2010s

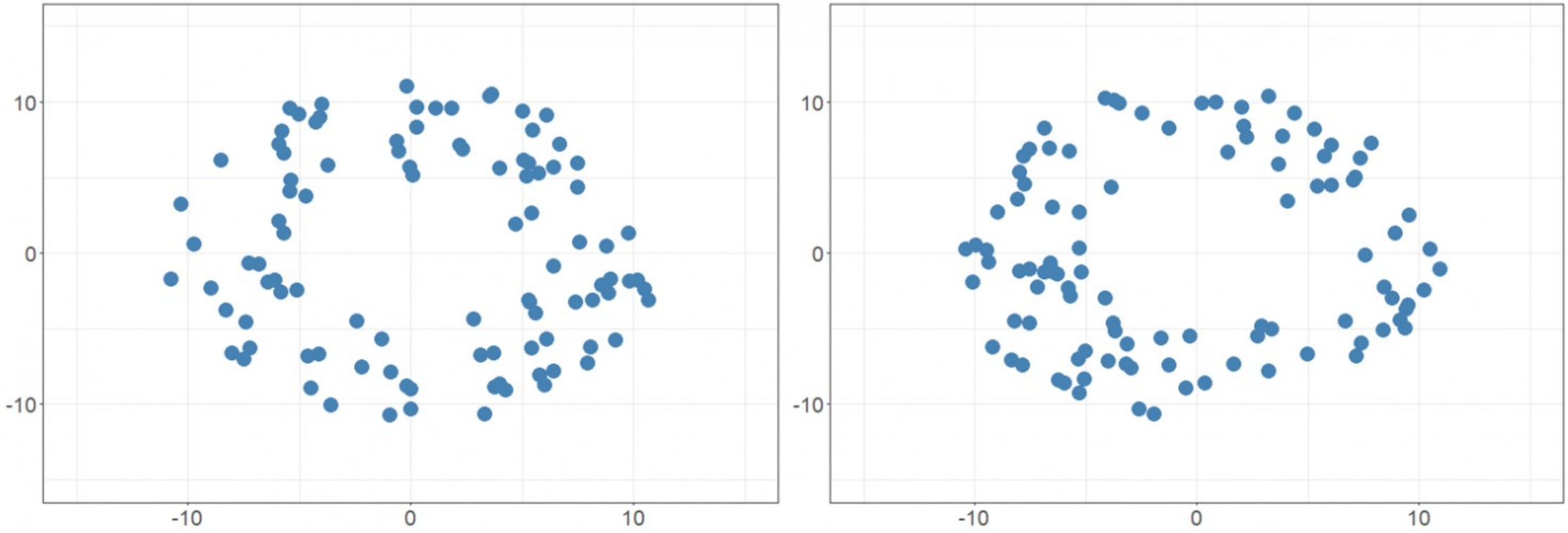
Example of the Progression in the Capabilities of GANs From 2014 to 2017. Taken from The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation, 2018.

StyleGAN can generate nearly arbitrary pose and clothing options.

Figure 9. **InsetGAN Results.** We show the combined results of six different human bodies generated from the given baseline model and six faces generated from the FFHQ [40] model. For every single face, shown on the left corner of each grid, we jointly optimize it with three different bodies.

https://stylegan-human.github.io/

GANs can also create synthetic data.



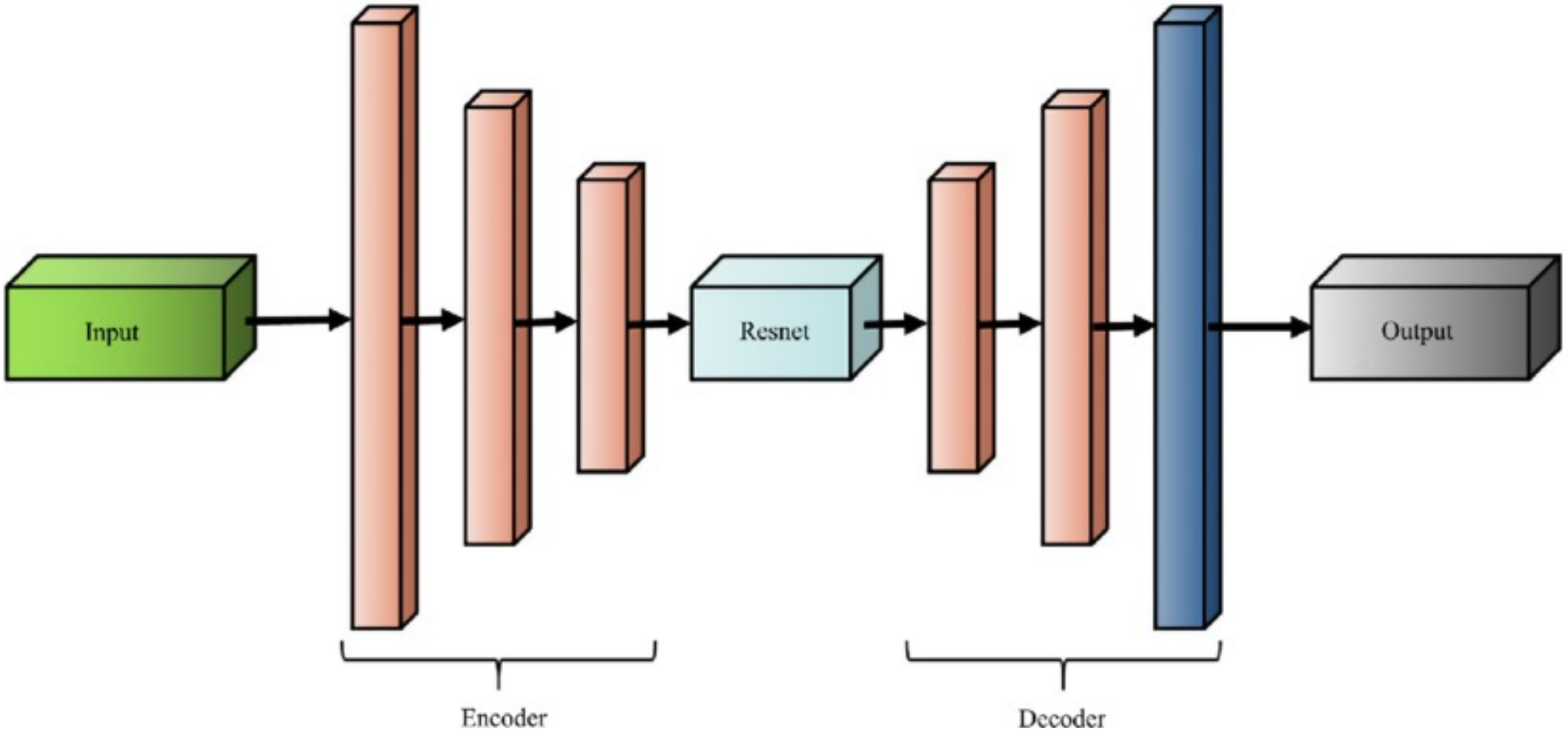
Original data

Synthetic data

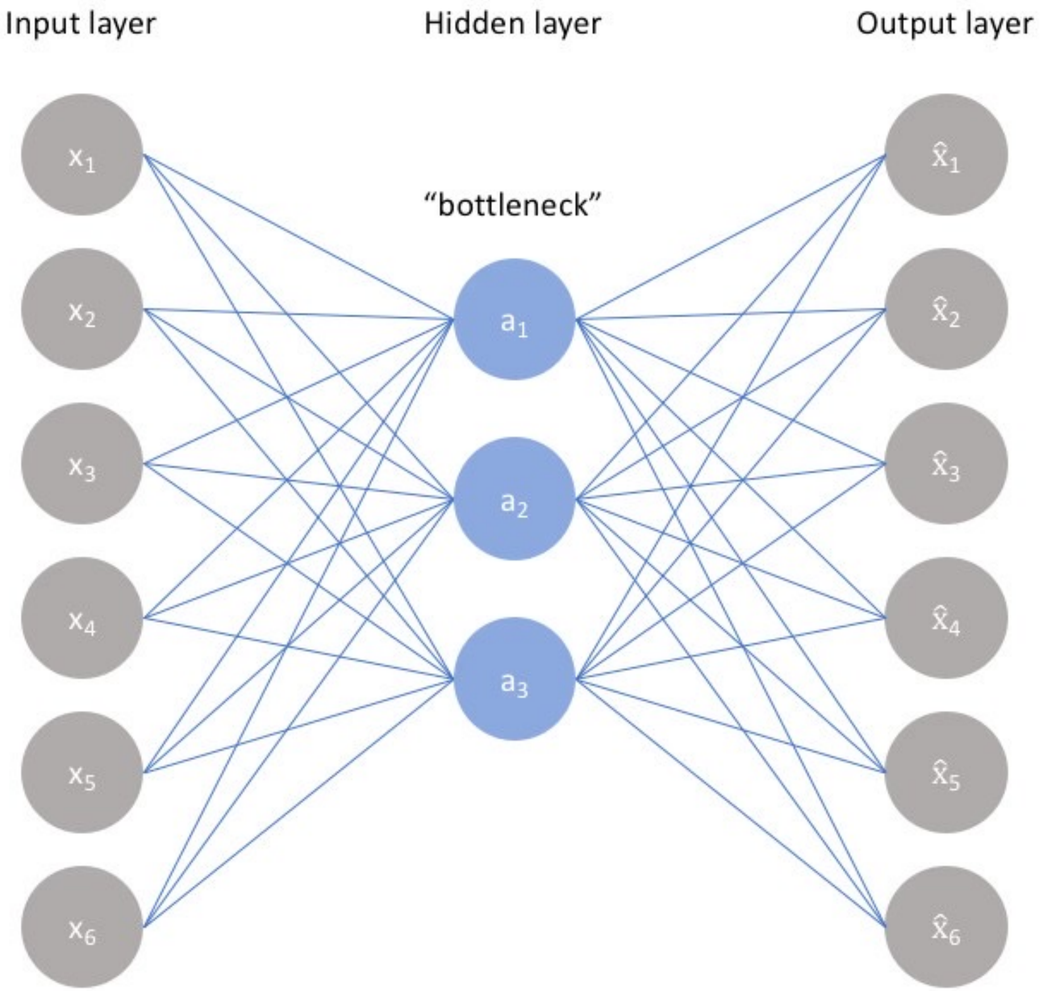
The synthetic data retains the structure of the original data but is not the same

https://dataingovernment.blog.gov.uk/2020/08/20/synthetic-data-unlocking-the-power-of-data-and-skills-for-machine-learning/

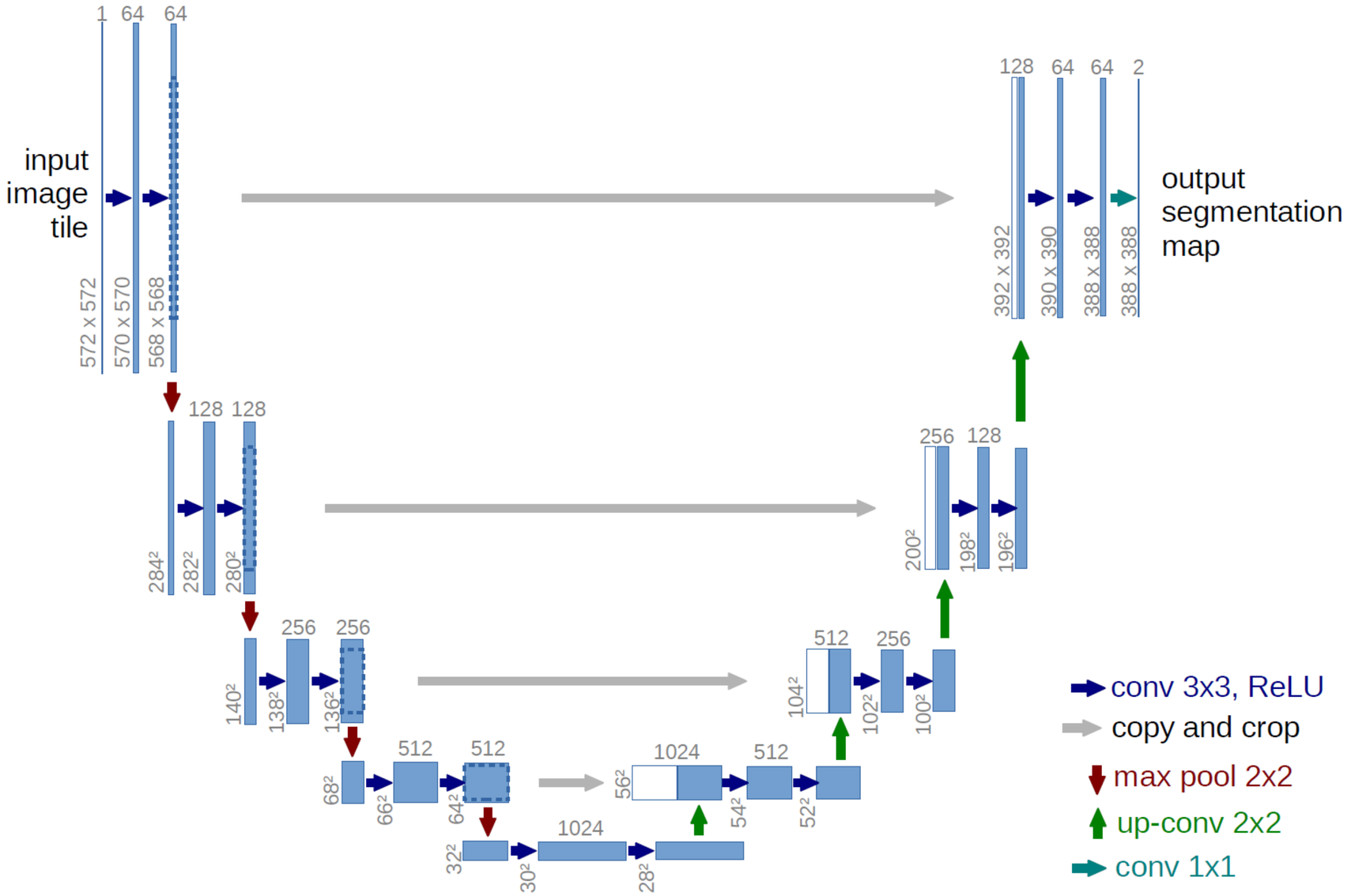
Encoder-Decoder Architectures



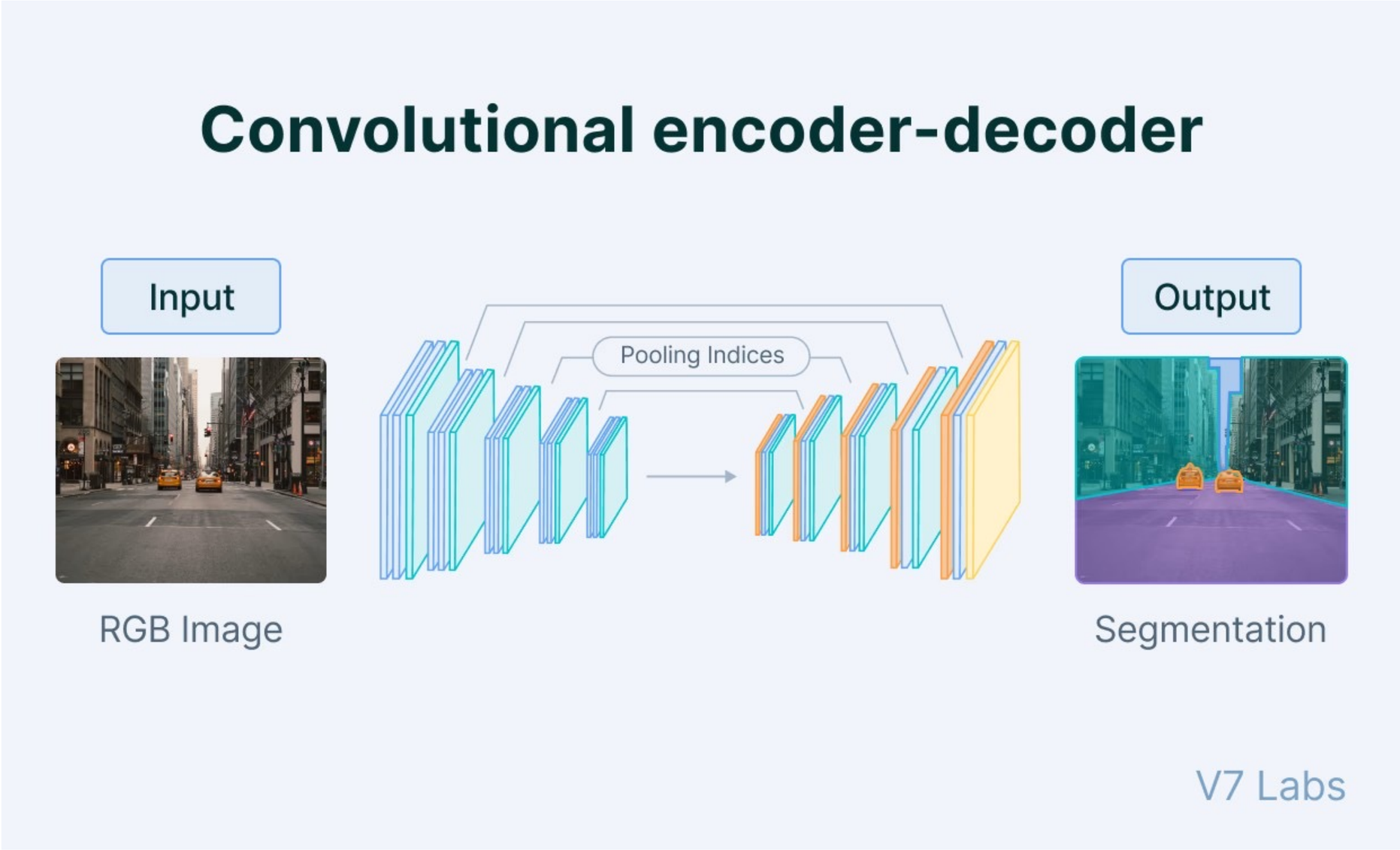
Autoencoders are trained neural nets that reconstruct an input from a more limited representation.



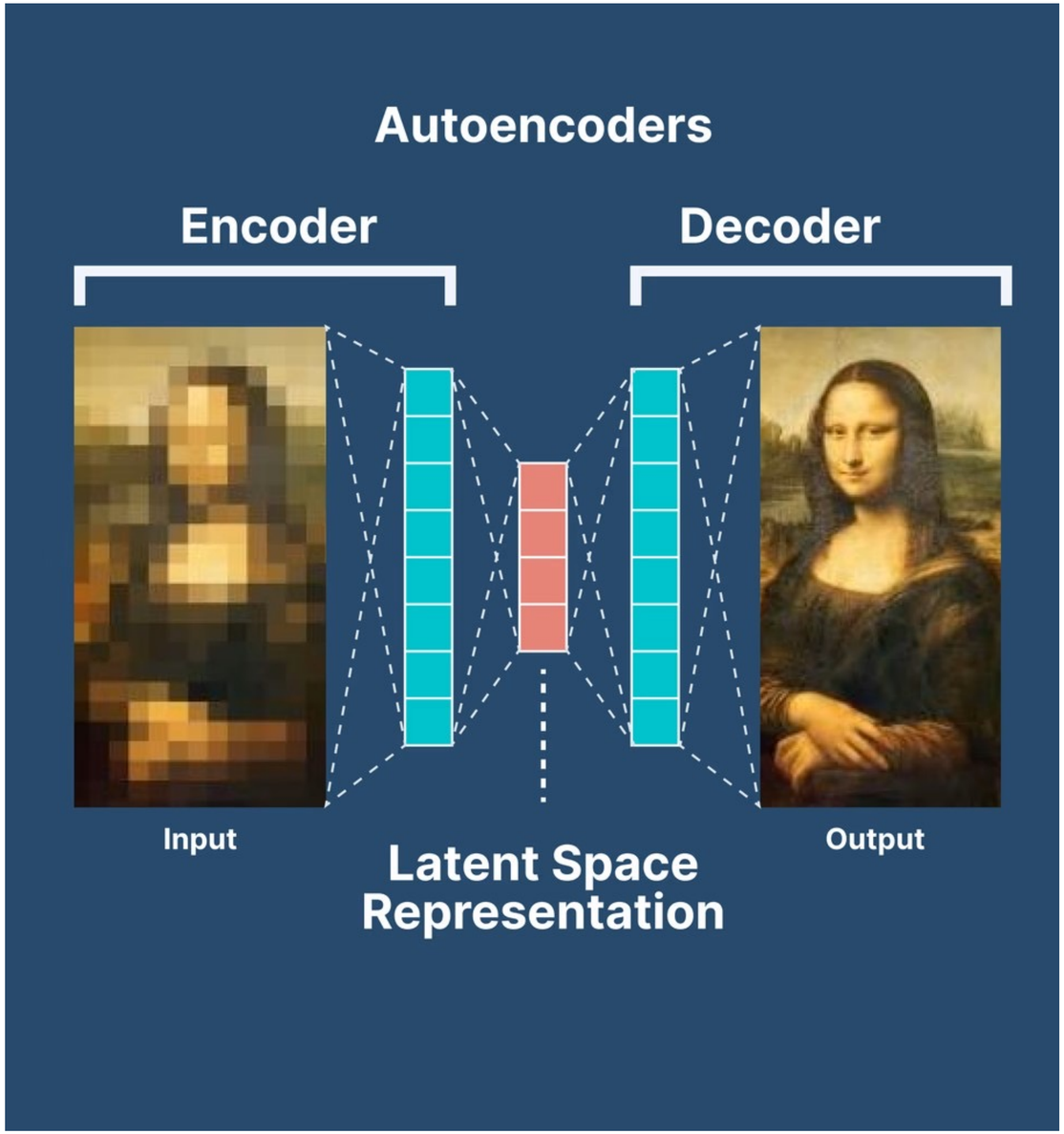
This architecture is *sometimes* called a “U-Net”



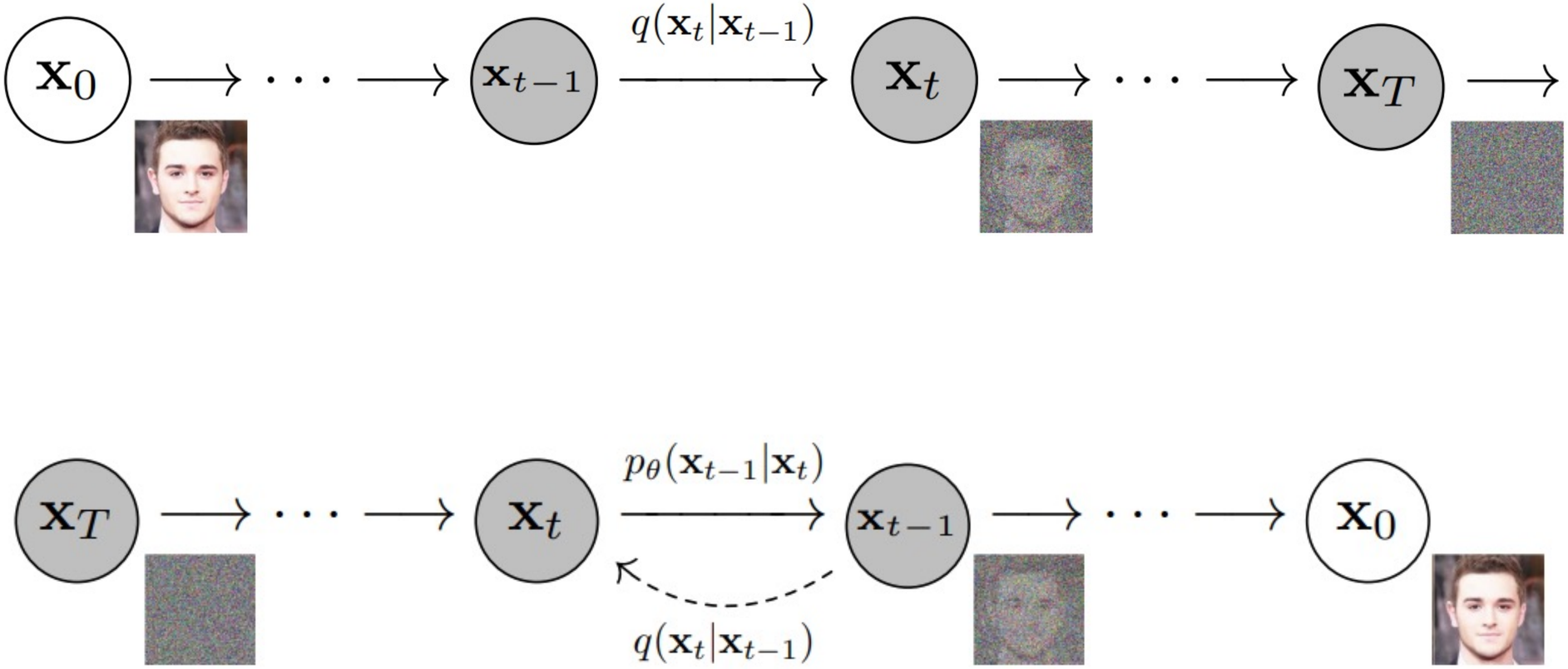
Autoencoders are analogous to CNNs used for image recognition, except they also reverse the process.



Autoencoders can denoise images.

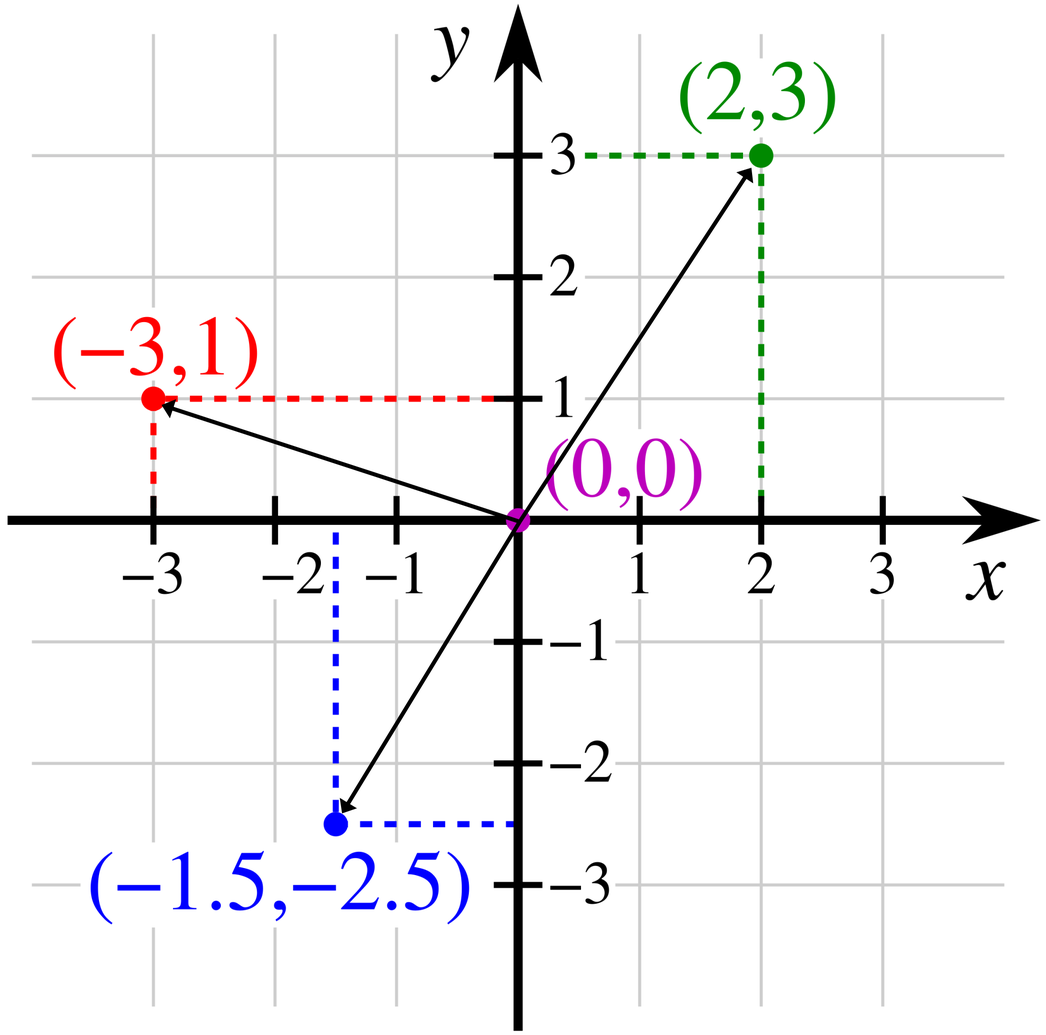


Diffusion models train to recover original images from noise.



A Short Interlude for Math

A vector is a list of numbers defining a point in a conceptual space.



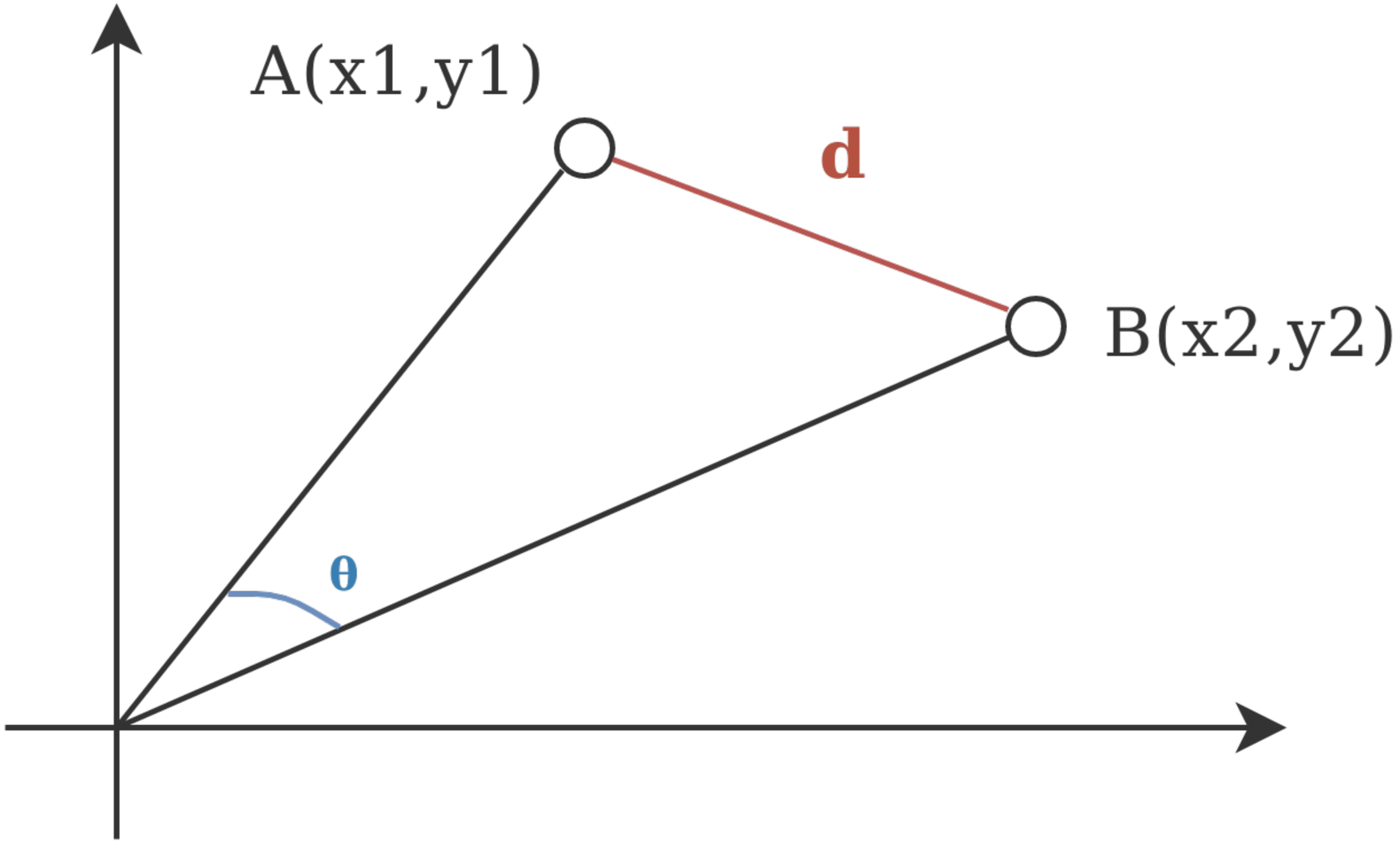
[2, 3]

[0, 0]

[-3, 1]

[-1.5, -2.5]

The difference between two vectors can be calculated by angle or by distance.



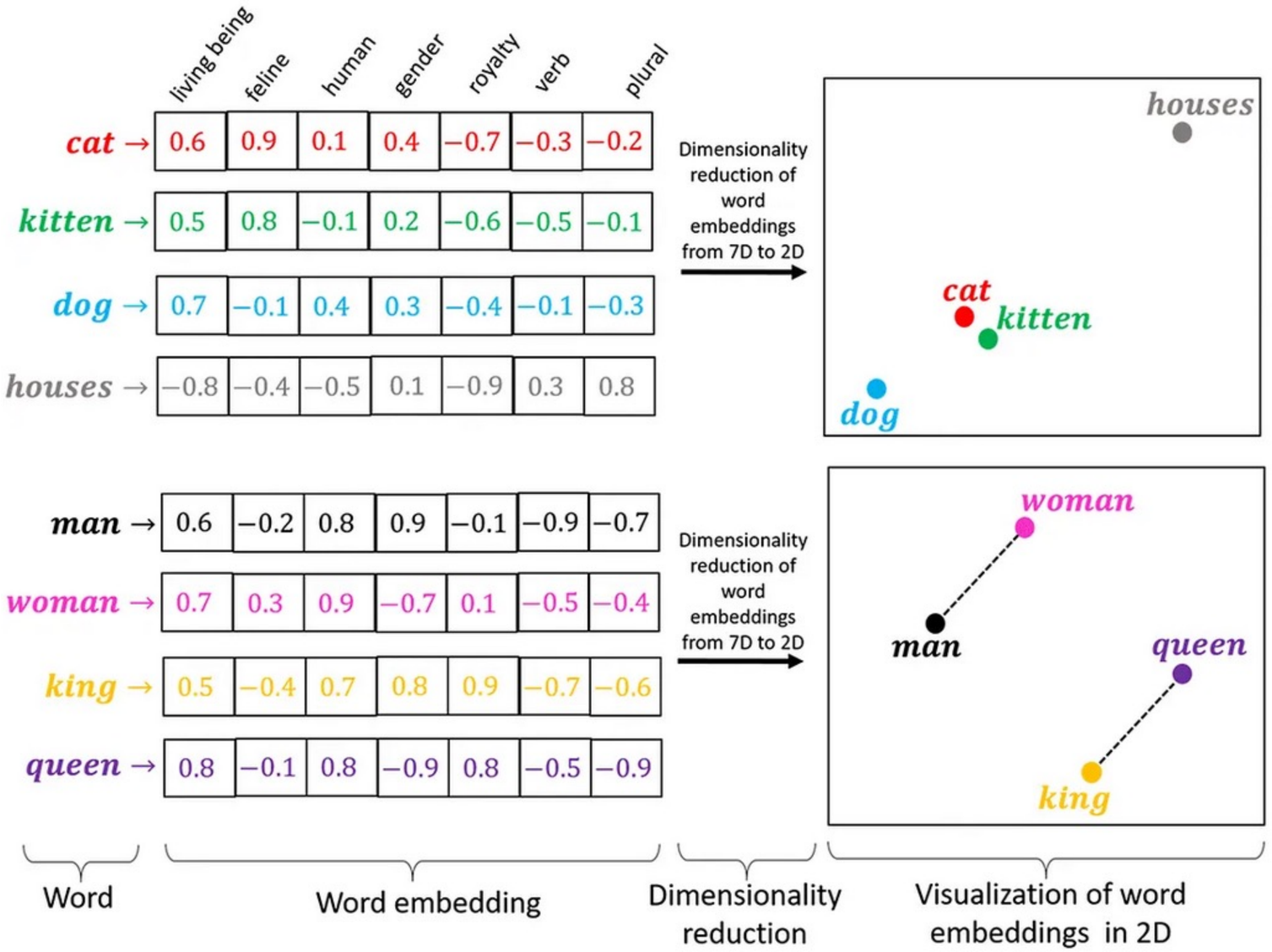
Terminology for vectors can be confusing – just recognize these words in context.

“vector” = “embedding” = “template” = a list of numbers

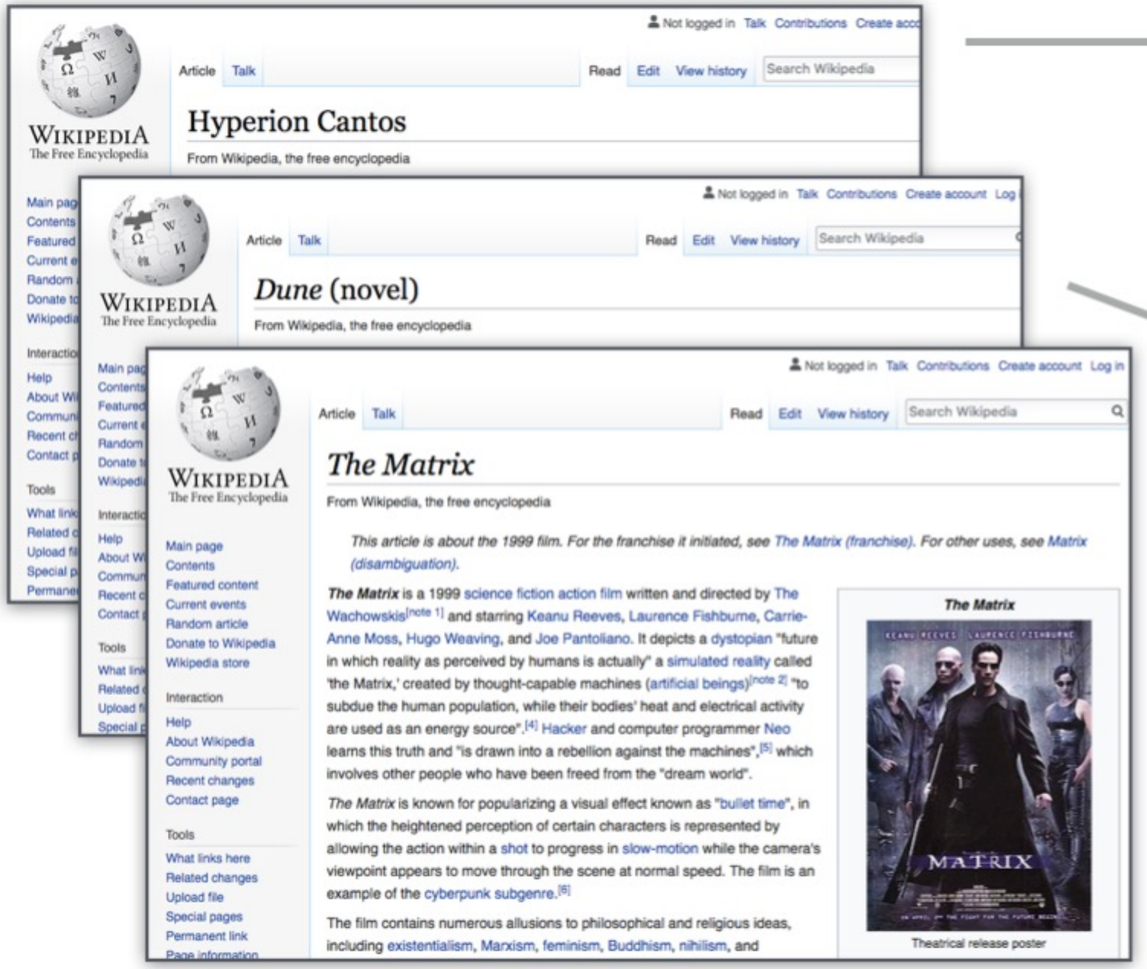
(One possible) embedding for 'embeddings': [-2.7053456 2.1384873 0.6417674
1.2882394 0.53470695 0.5651745 0.64166373 -1.1691749 0.32658175 -
0.99961764]

Word Embeddings

Word embeddings (vectors) represent how similar two words are.



Word embeddings are (usually) generated by their relationship to other words in documents.



The **Hyperion Cantos** is a series of science fiction novels by Dan Simmons. The title in the series, *Hyperion* and *The Fall of Hyperion*,[1][2] and later came to refer to the overall storyline, including *Endymion*, *The Rise of Endymion*, and a number of short stories.[3][4] More narrowly, inside the fictional storyline, after the first volume, the Hyperion Cantos is an epic poem written by the character Martin Silenus covering in verse form the events of the first book.[5]

Of the four novels, *Hyperion* received the Hugo and Locus Awards in 1990;[6] *The Fall of Hyperion* won the Locus and British Science Fiction Association Awards in 1991;[7] and *The Rise of Endymion* received the Locus Award in 1998.[8] All four novels were also nominated for various science fiction awards.

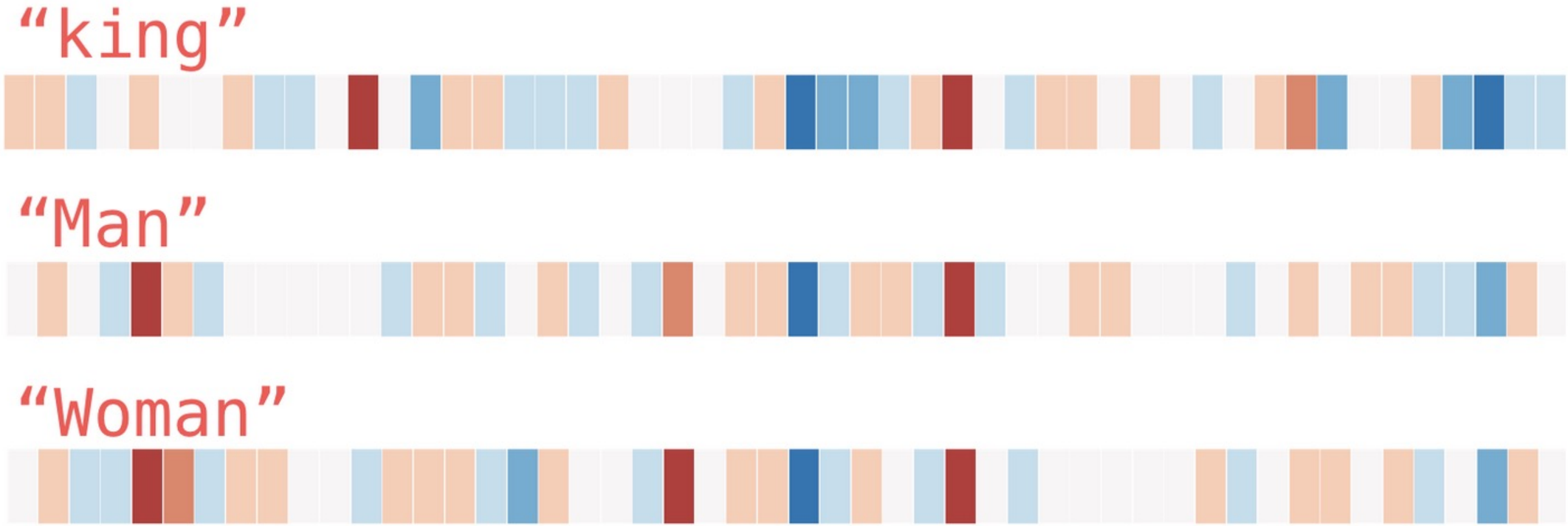
An event series is being developed by Bradley Cooper, Graham King, and Todd Phillips for Syfy based on the first novel *Hyperion*.[9]

Dune is a 1965 science fiction novel by American author Frank Herbert, originally published as two separate serials in *Analog* magazine. It tied with Roger Zelazny's *This Immortal* for the Hugo Award in 1966,[3] and it won the inaugural Nebula Award for Best Novel.[4] It is the first installment of the *Dune* saga, and in 2003 was cited as the world's best-selling science fiction novel.[5][6]

Set in the distant future amidst a feudal interstellar society in which noble houses, in control of individual planets, owe allegiance to the Padishah Emperor, *Dune* tells the story of young Paul Atreides, whose noble family accepts the stewardship of the planet Arrakis. It is an inhospitable and sparsely populated desert wasteland, but is also the only source of melange, also known as "spice", a drug that enhances mental abilities. As melange is the most important and valuable substance in the universe, control of Arrakis is a coveted—and dangerous—undertaking. The story explores the multi-layered interactions of politics, religion, ecology, technology, and human emotion, as the factions of the empire confront each other in a struggle for the control of Arrakis

The Matrix is a 1999 science fiction action film written and directed by The Wachowskis[note 1] and starring Keanu Reeves, Laurence Fishburne, Carrie-Anne Moss, Hugo Weaving, and Joe Pantoliano. It depicts a dystopian "future in which reality as perceived by humans is actually" a simulated reality called 'the Matrix,' created by thought-capable machines (artificial beings)[note 2] "to subdue the human population, while their bodies' heat and electrical activity

Word embeddings show patterns of similarity for similar or related words.



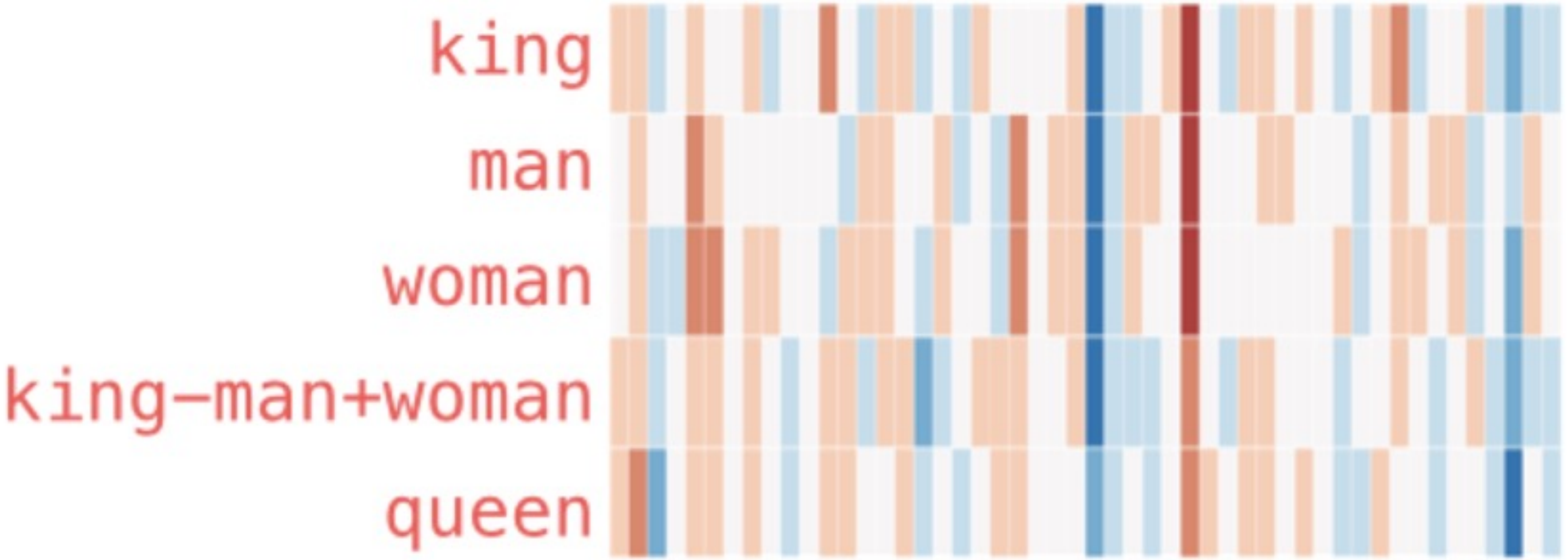
Semantle lets you play “guess the word” using word embedding distance.

#	Guess	Similarity	Proximity
64	counter	100.00	🔥🎉
Sort – Similarity			
38	table	30.21	940 / 1000
63	sandwich	24.70	668 / 1000
23	kitchen	24.44	639 / 1000
43	bar	23.99	575 / 1000
33	plate	23.40	490 / 1000
29	restaurant	22.29	258 / 1000
52	menu	21.91	170 / 1000
21	refrigerator	21.24	(tepid)
31	diner	20.97	(tepid)
37	glass	19.04	(cold)

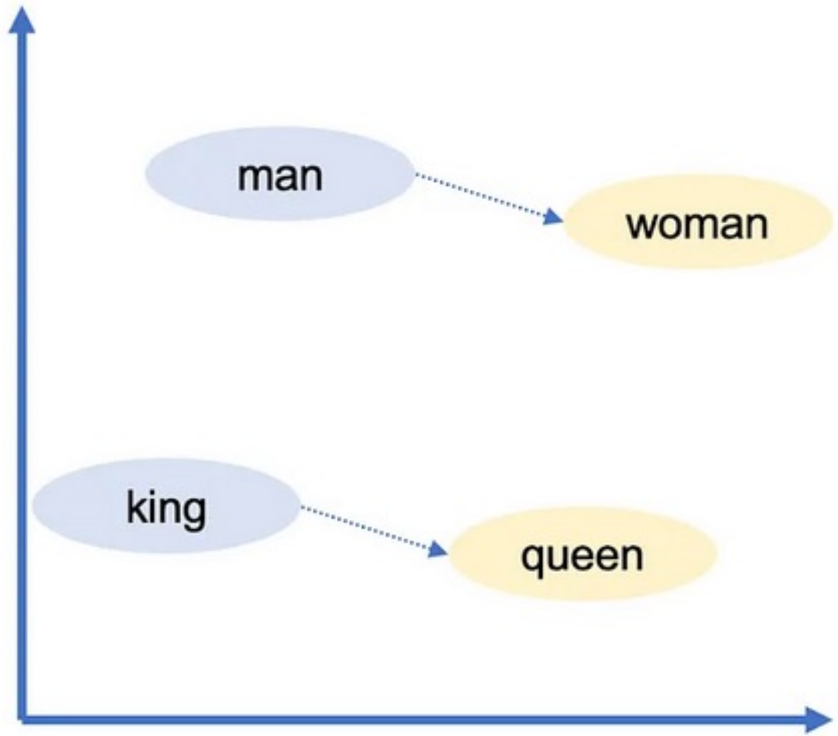
https://semantle.com/

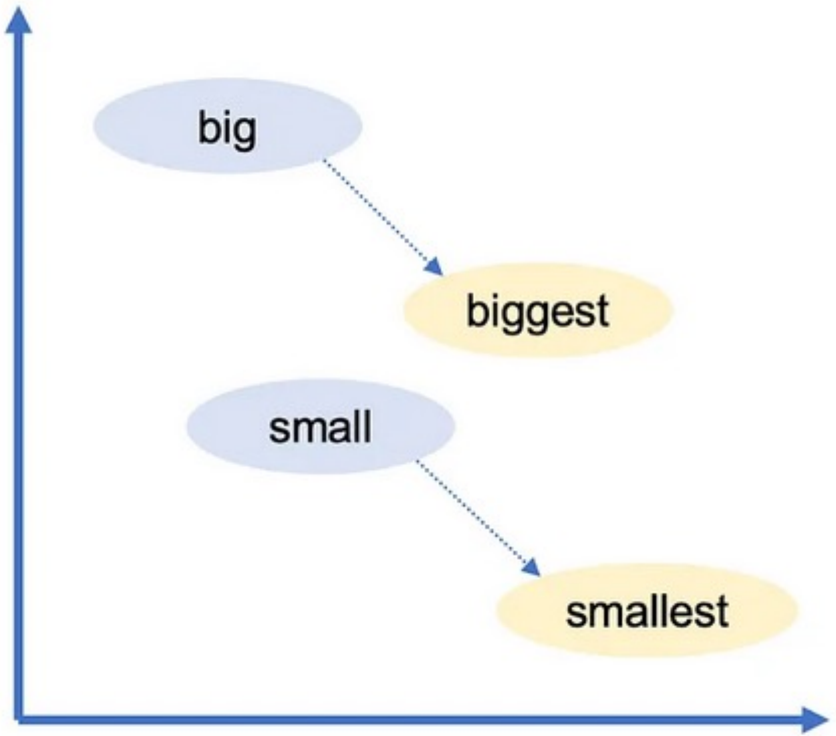
You can do math on word embeddings.

king − man + woman ~= queen



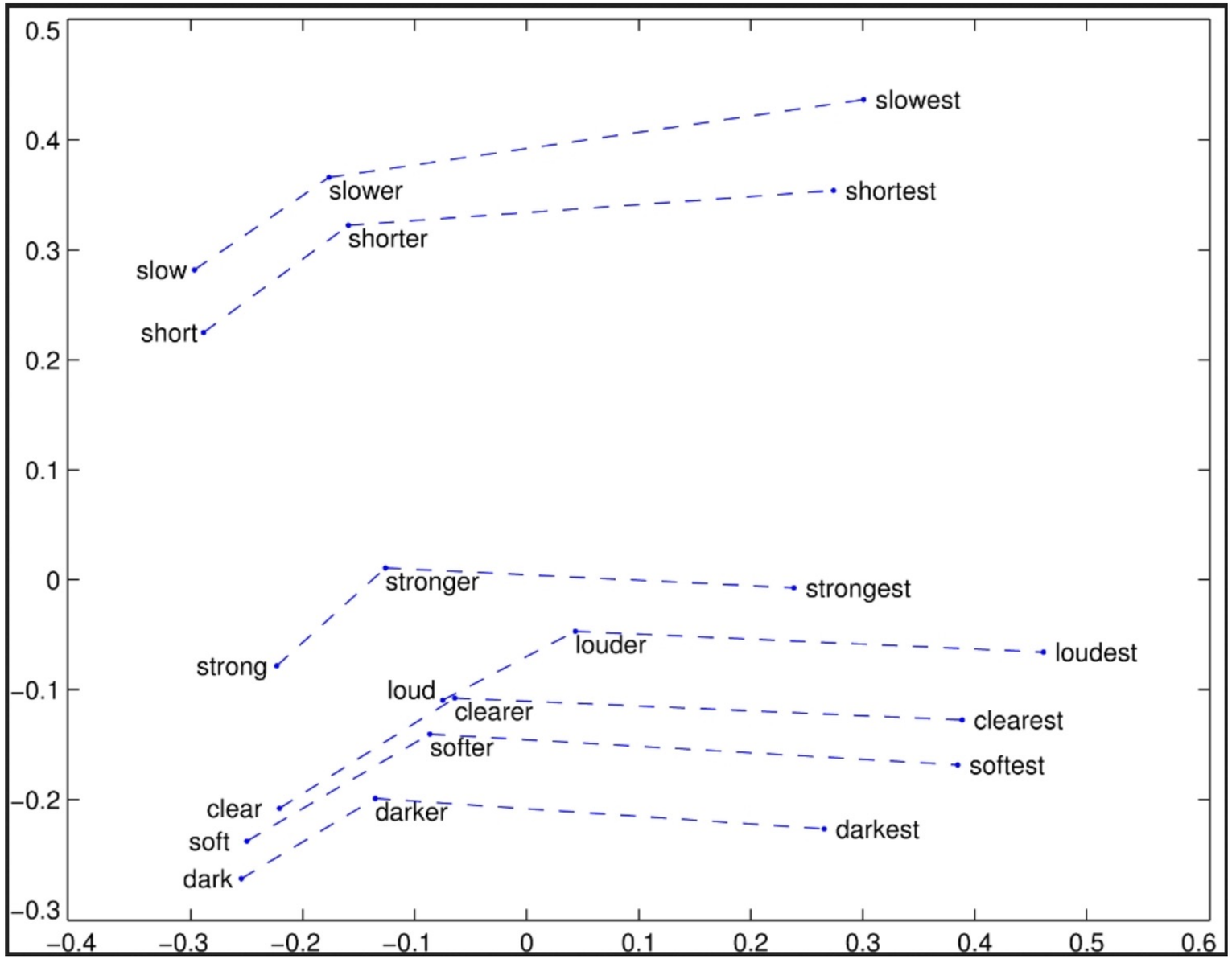
(Artificially dimension-reduced) visualization of word embedding math.





https://www.understandingai.org/p/large-language-models-explained-with

More examples of embedding relations.



Transformers

Attention is All You Need, 2017

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

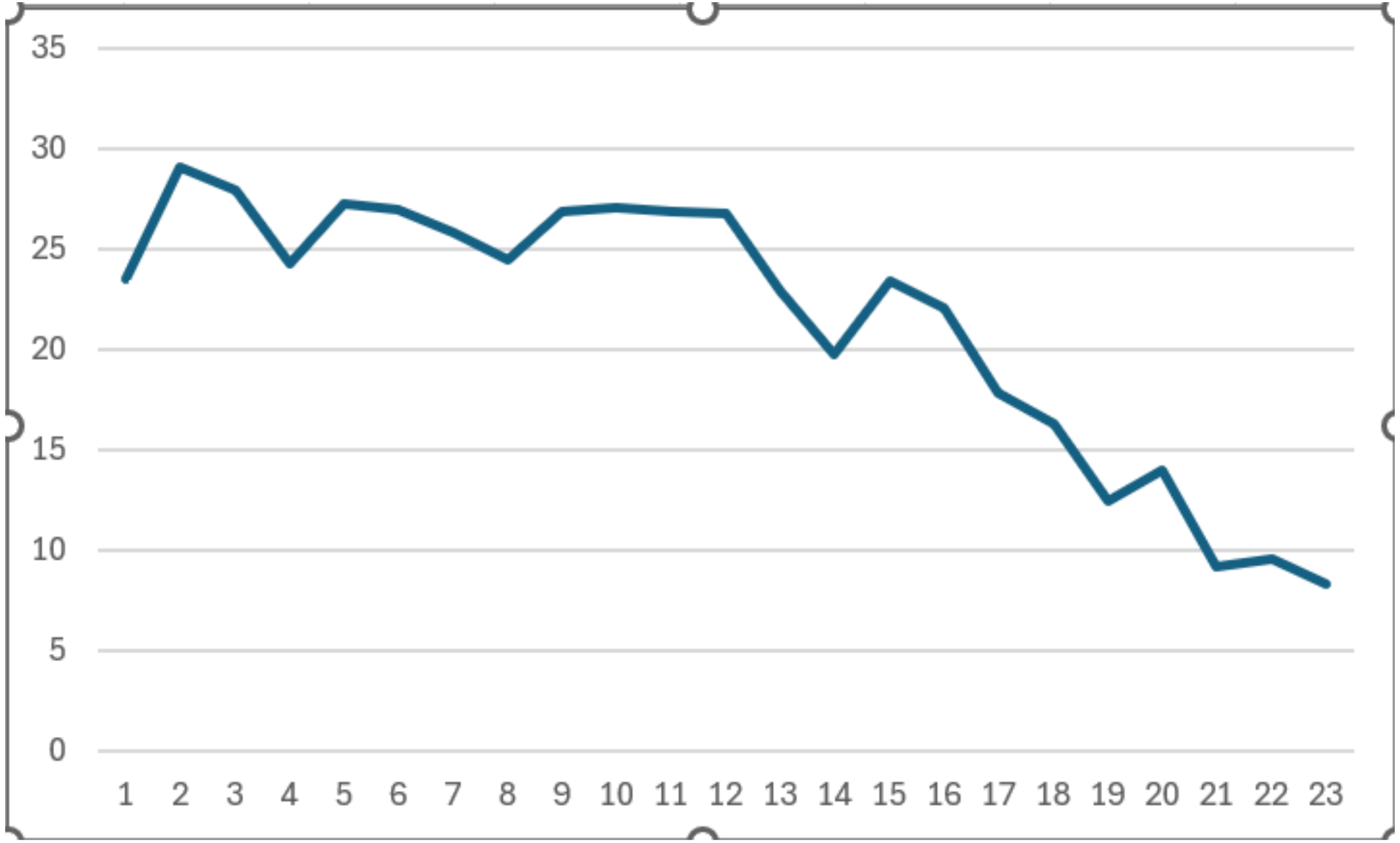
Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

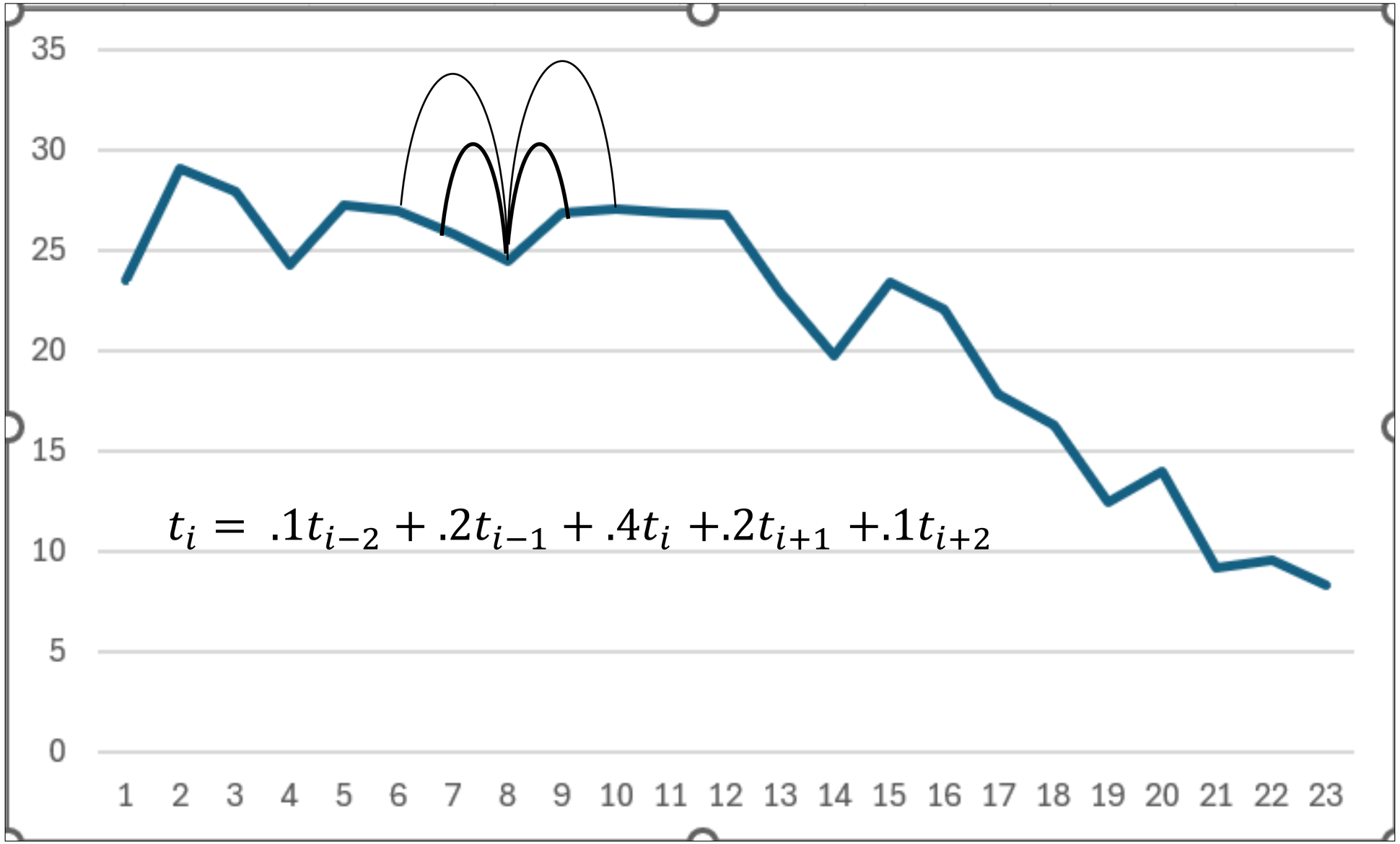
Abstract

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.

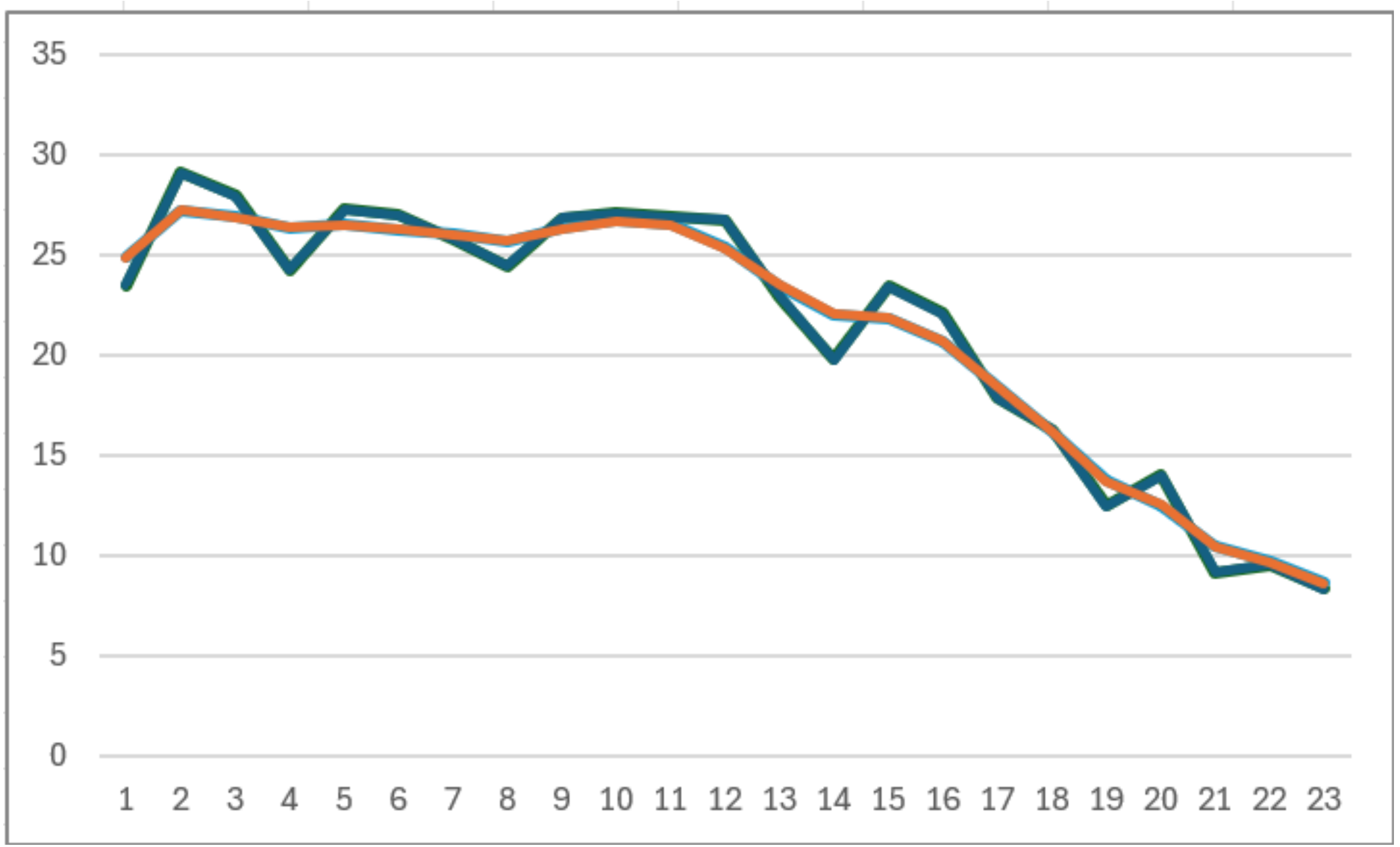
Motivation: sometimes data are noisy and we would like to get a cleaner signal.



We could weight data points by their neighbors to improve the signal.



Each point has been updated with context from its neighbors.



We want to predict the next word in a sequence.

“The giant anteater can be identified by its large size, long, narrow muzzle, and long, bushy…”

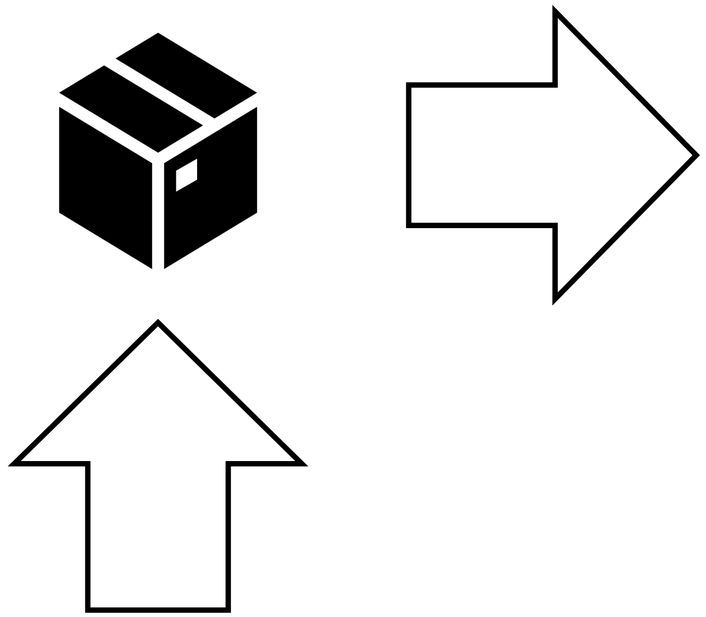
https://en.wikipedia.org/wiki/Giant_anteater

What ChatGPT told me

Based on the sentence you provided, here are some possible words that could follow "long, bushy," along with estimated probabilities for each guess:

1. **tail** (30%)
    - The phrase "long, bushy tail" is commonly used to describe animals, particularly when referencing their physical features.
2. **fur** (20%)
    - Anteaters are often described in terms of their fur or coat, and "long, bushy fur" fits the pattern.
3. **ears** (15%)
    - Animals like anteaters are sometimes described with focus on their long ears, especially when characterized as large and noticeable.
4. **coat** (10%)
    - "Long, bushy coat" is another possibility, since animals with prominent fur features might be described this way.

We’d like to use the word embeddings as an input to a predicting neural net.



tail (.3)
fur (.2)
ears (.15)
coat (.1)
hair (.08)
whiskers (.07)
beard (.005)
plumes (.003)

“The giant anteater can be identified by its large size, long, narrow muzzle, and long, bushy…”

Can we improve the quality of the “signal” here by providing more context?

v_1 v_2 v_3 v_4 v_5 v_6 v_7 v_8 v_9

“giant anteater can be identified by its long, bushy…”

Quick math review on multiplying vectors.

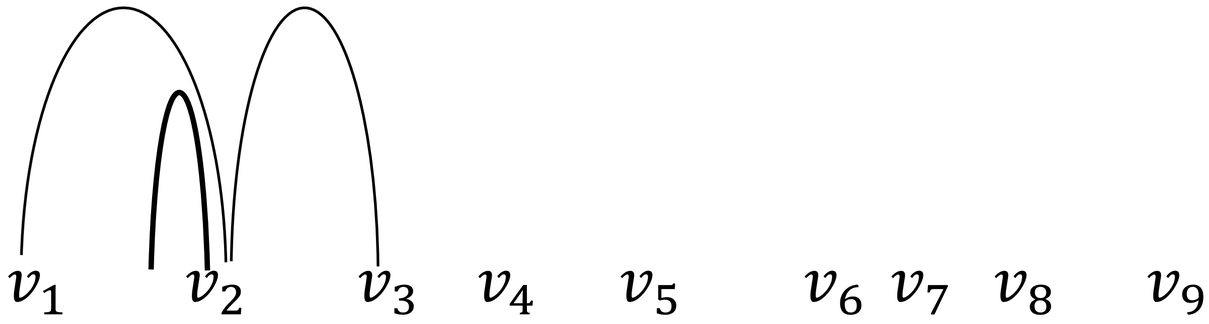
Dot product:

$$[a_1 \quad b_1 \quad c_1] \cdot [a_2 \quad b_2 \quad c_2] = a_1\, a_2 + b_1 b_2 + c_1 c_2$$
$$= \cos\theta |a|\, |b|$$

What if we weigh each vector (word embedding) by the dot product of each other word?

$$s_{12} = v_1 \cdot v_2$$

$$s_{22} = v_2 \cdot v_2$$



"giant anteater can be identified by its long, bushy…"

What if we weigh each vector (word embedding) by the dot product of each other word?

$$s_{12} = v_1 \cdot v_2$$
$$s_{22} = v_2 \cdot v_2$$
$$s_{32} = v_3 \cdot v_2$$
$$s_{42} = v_4 \cdot v_2$$
$$s_{52} = v_5 \cdot v_2$$
$$s_{62} = v_6 \cdot v_2$$
$$s_{72} = v_7 \cdot v_2$$
$$s_{82} = v_8 \cdot v_2$$
$$s_{92} = v_9 \cdot v_2$$

$$\sum_i w_{ij} = 1$$

→ Normalize →

$$w_{12}$$
$$w_{22}$$
$$w_{32}$$
$$w_{42}$$
$$w_{52}$$
$$w_{62}$$
$$w_{72}$$
$$w_{82}$$
$$w_{92}$$

$$y_2 = w_{12} \cdot v_2 + \ldots + w_{92} \cdot v_2$$

v_1 v_2 v_3 v_4 v_5 v_6 v_7 v_8 v_9

“giant anteater can be identified by its long, bushy…”

Some intuition on why attention might be helpful

Account

Bank

River

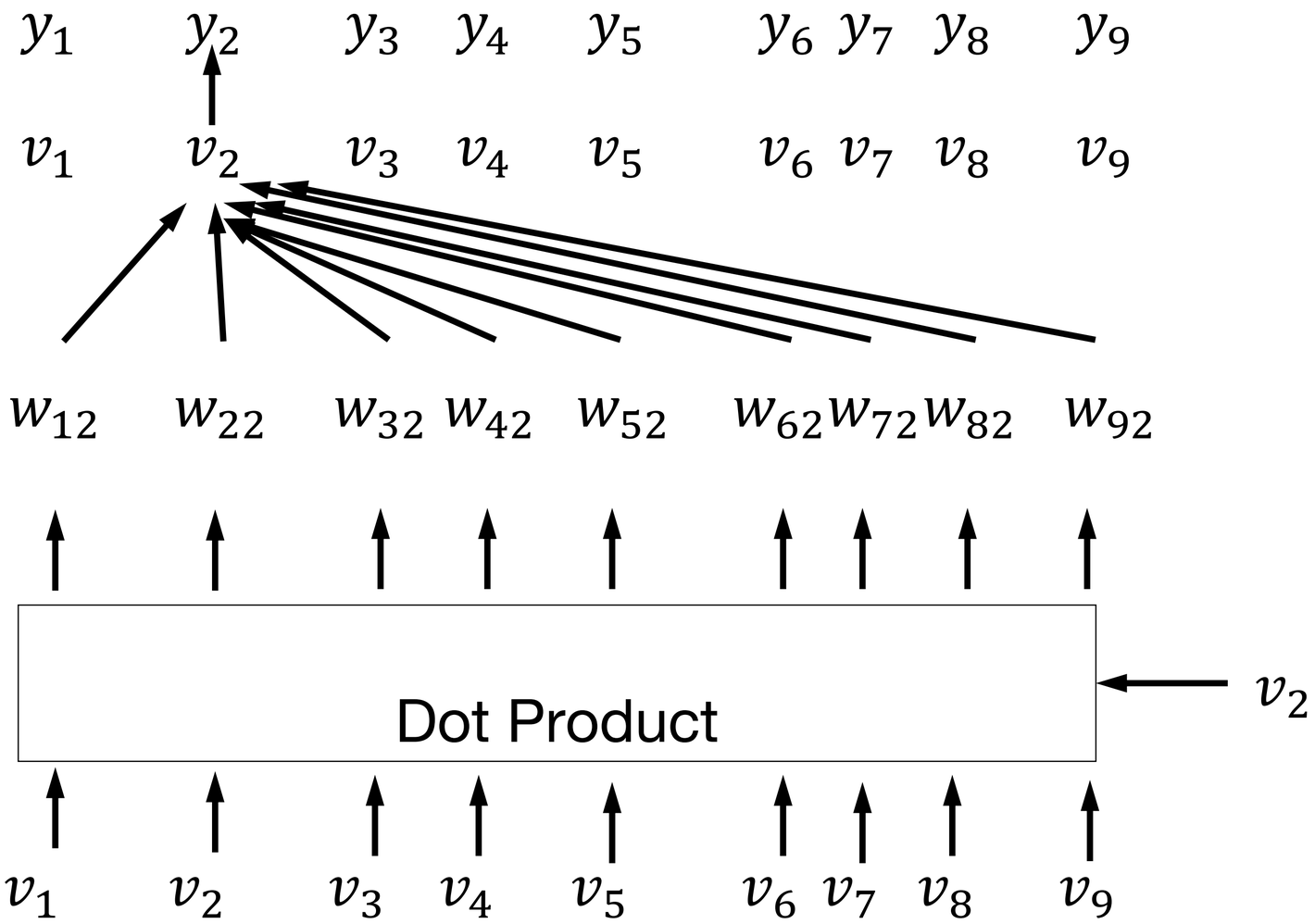
Sentence 1: John checked his account at the bank.

Sentence 2: John went down to the river bank.

Bank(1) ~= .5*bank + .25*account + …

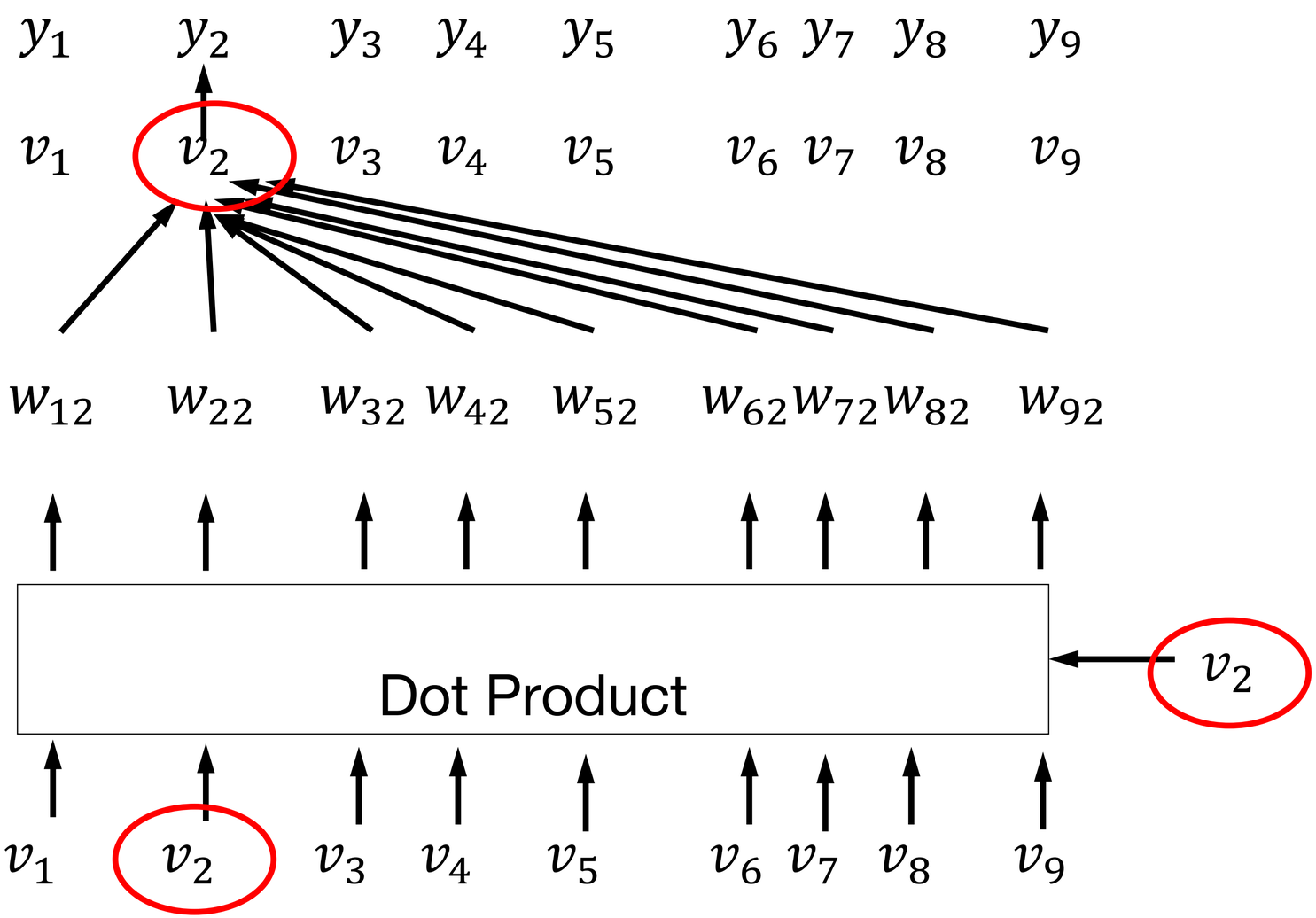
Bank(2) ~= .5*bank + .25*river + …

This process of re-weighing by each word is known as **self-attention**.



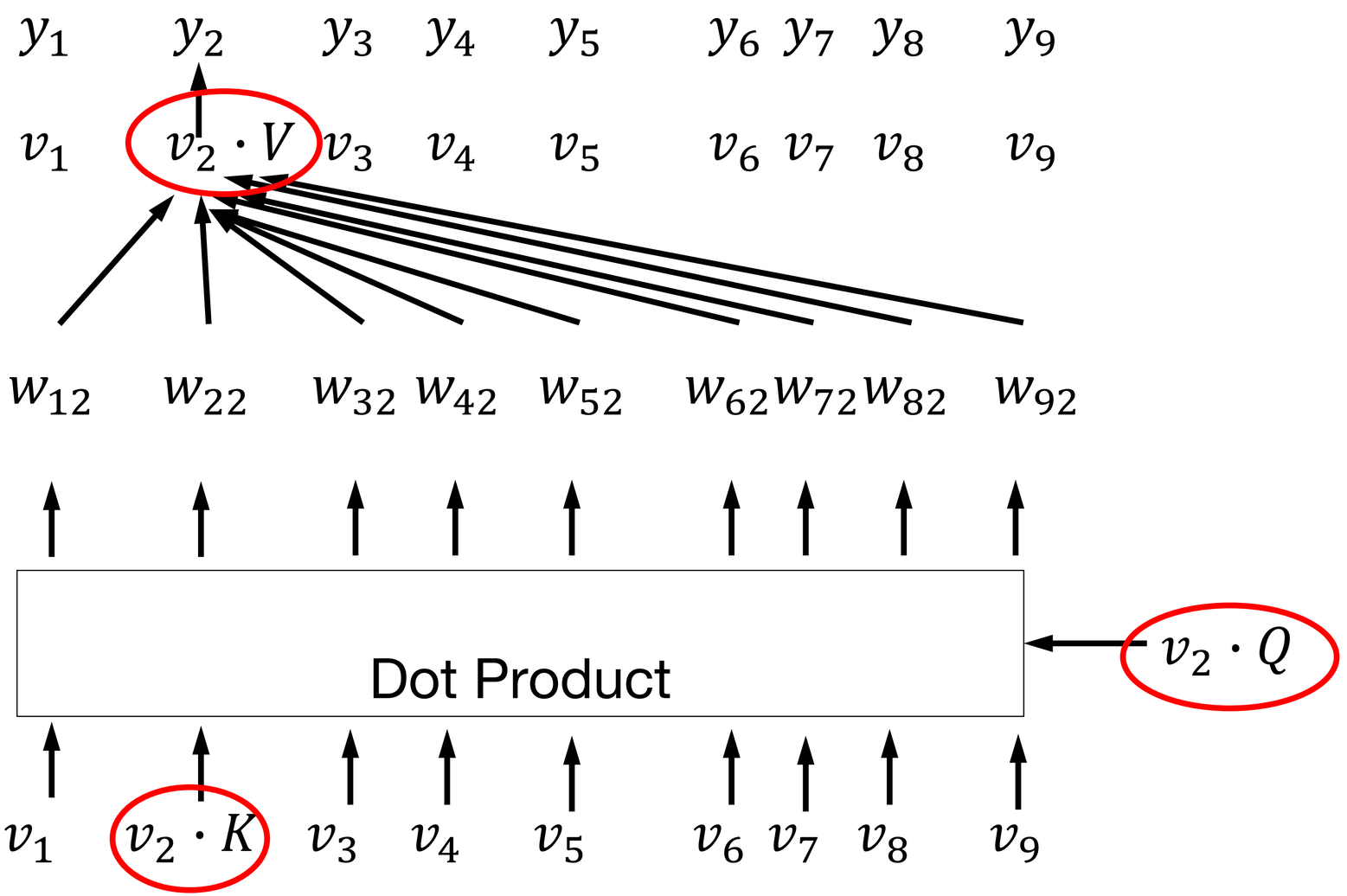
"giant anteater can be identified by its long, bushy…"

Note that the word we're applying attention to shows up three times in the process.



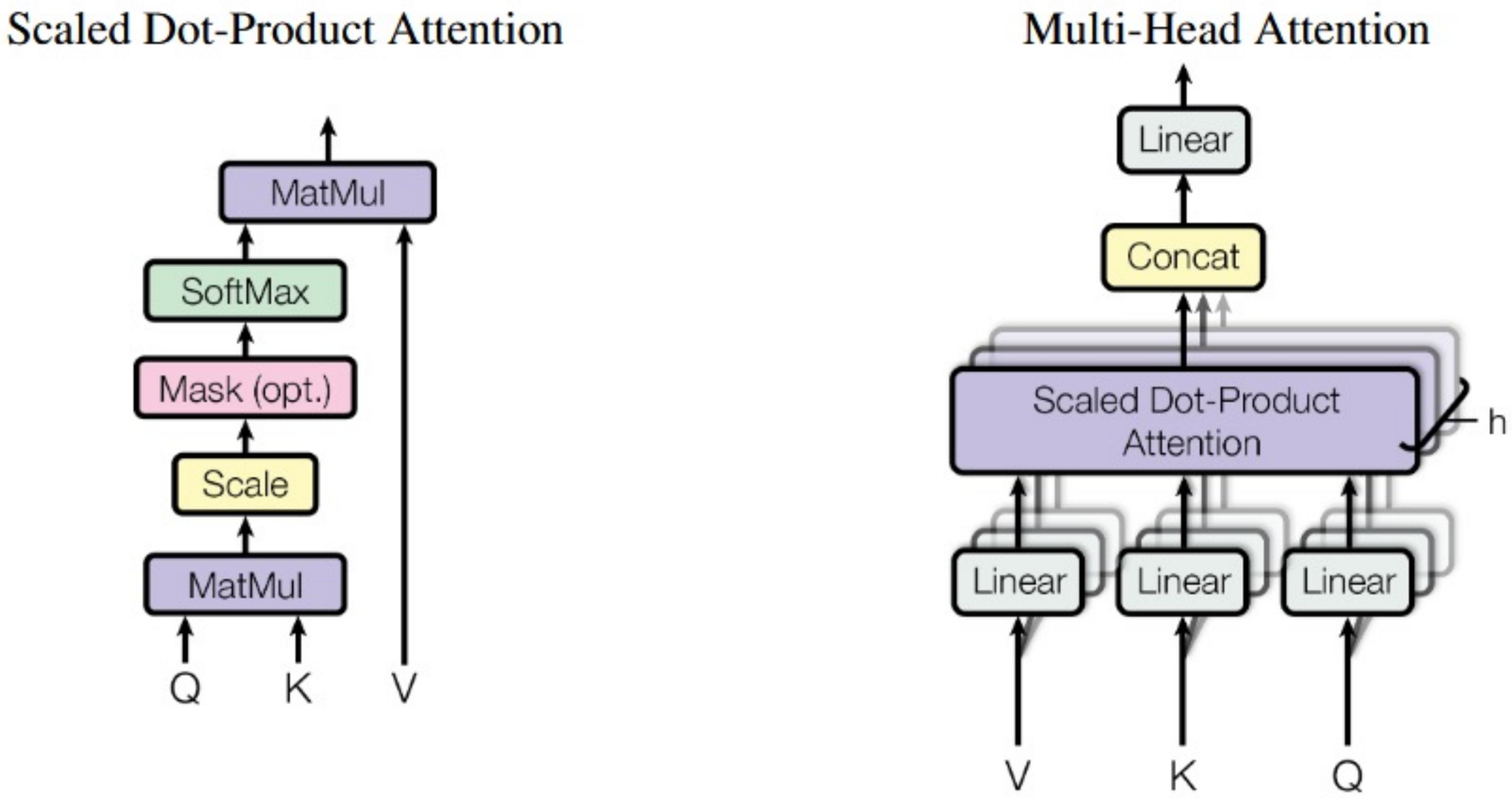
"giant anteater can be identified by its long, bushy…"

We weigh each use of the embedding by a trainable matrix (“Query,” “Key,” or “Value”).

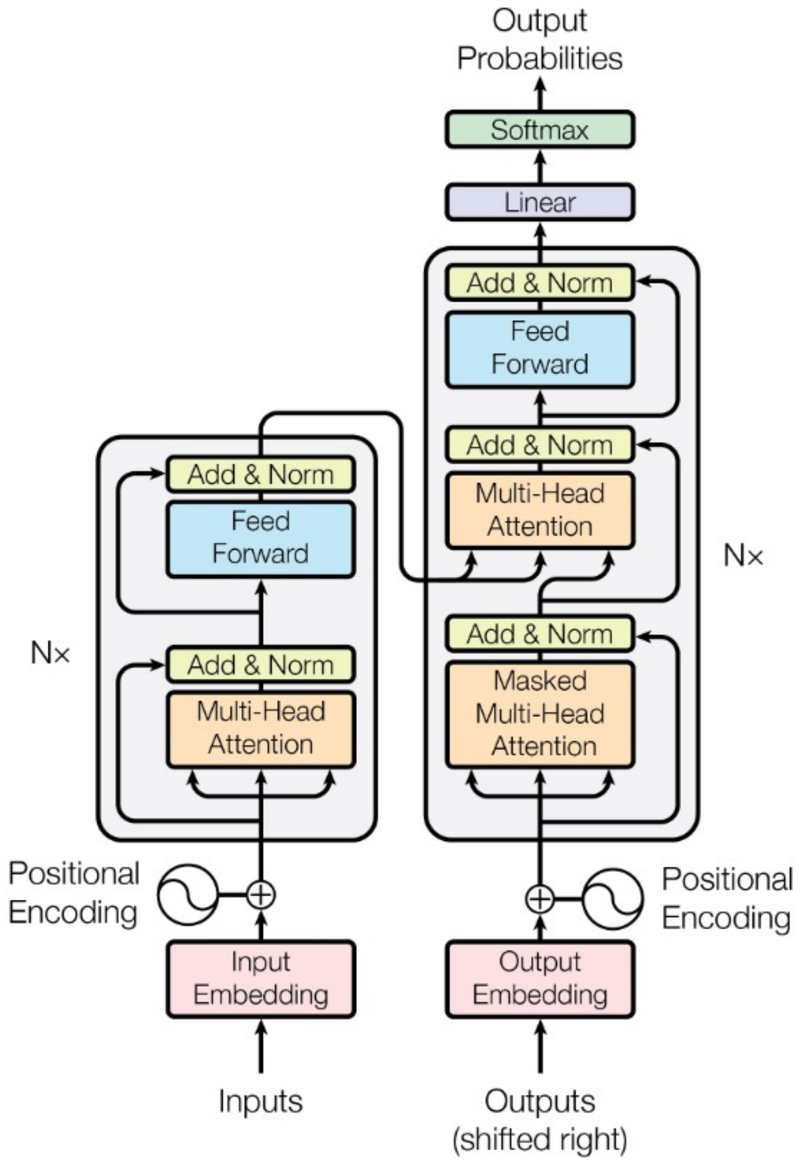


“giant anteater can be identified by its long, bushy…”

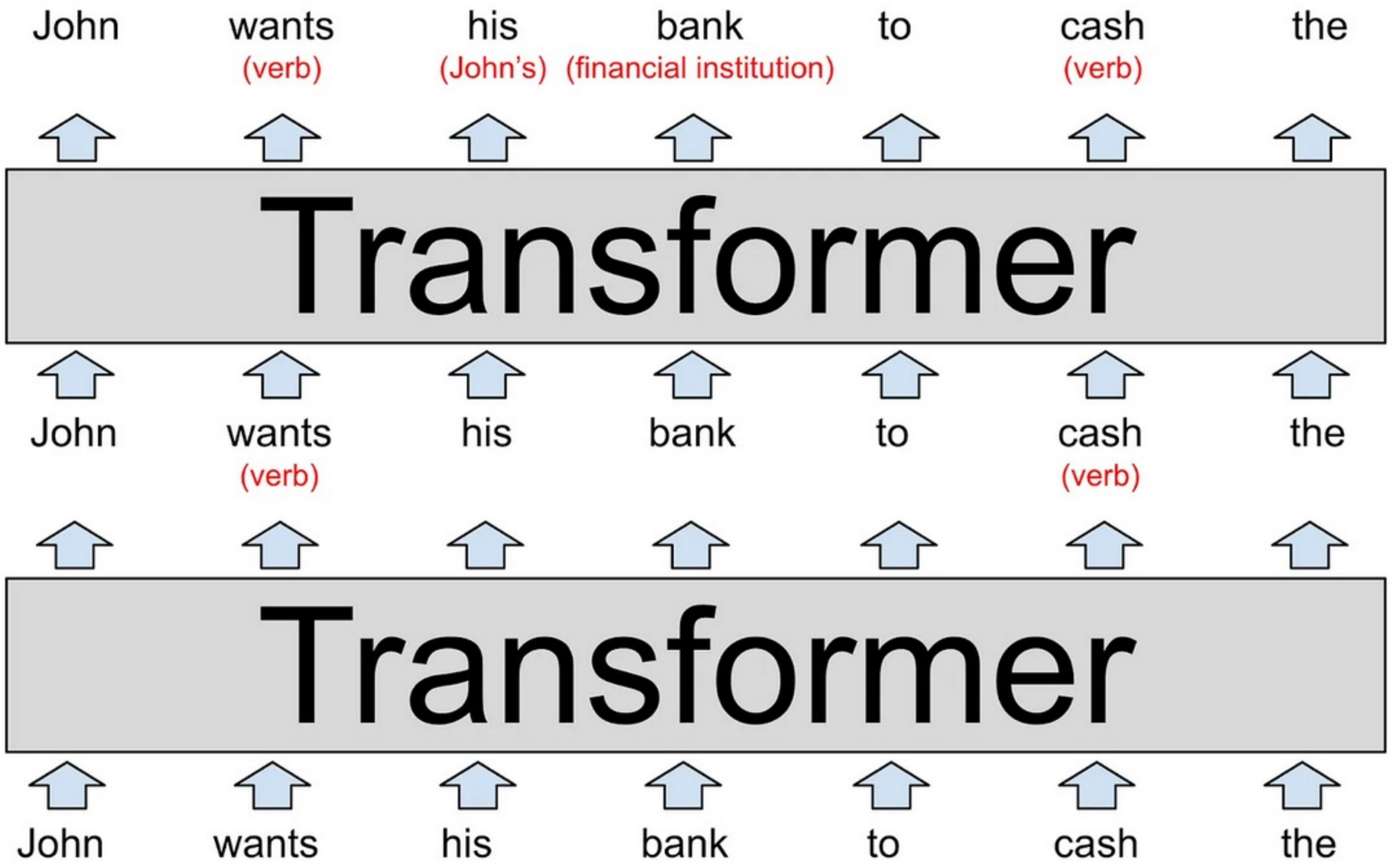
Attention

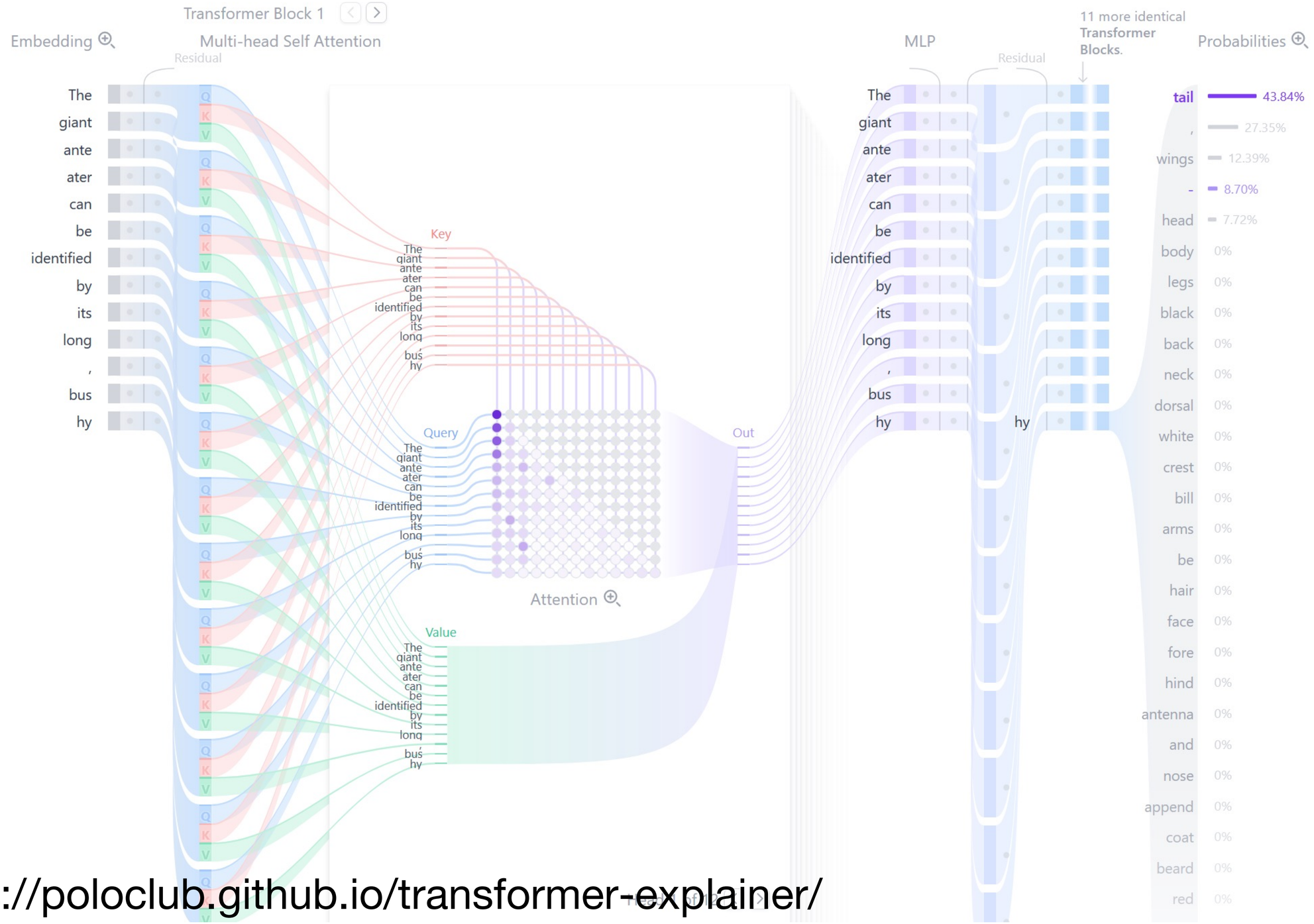

Scaled Dot-Product Attention
MatMul
SoftMax
Mask (opt.)
Scale
MatMul
Q
K
V
Multi-Head Attention
Linear
Concat
Scaled Dot-Product Attention
h
Linear
Linear
Linear
V
K
Q

The Transformer



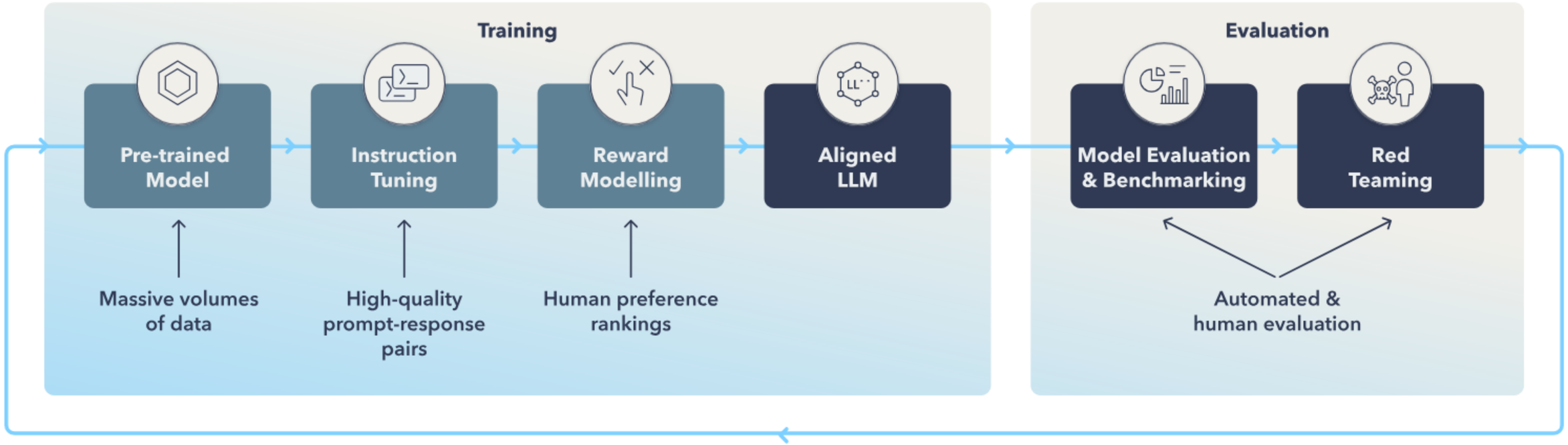
Multiple transformers each continue to add more context and relationships to the vector.





https://poloclub.github.io/transformer-explainer/

Putting a Large Language Model together



Does an LLM “think”? One point of view: They are “Stochastic Parrots.”

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜

Emily M. Bender*
ebender@uw.edu
University of Washington
Seattle, WA, USA

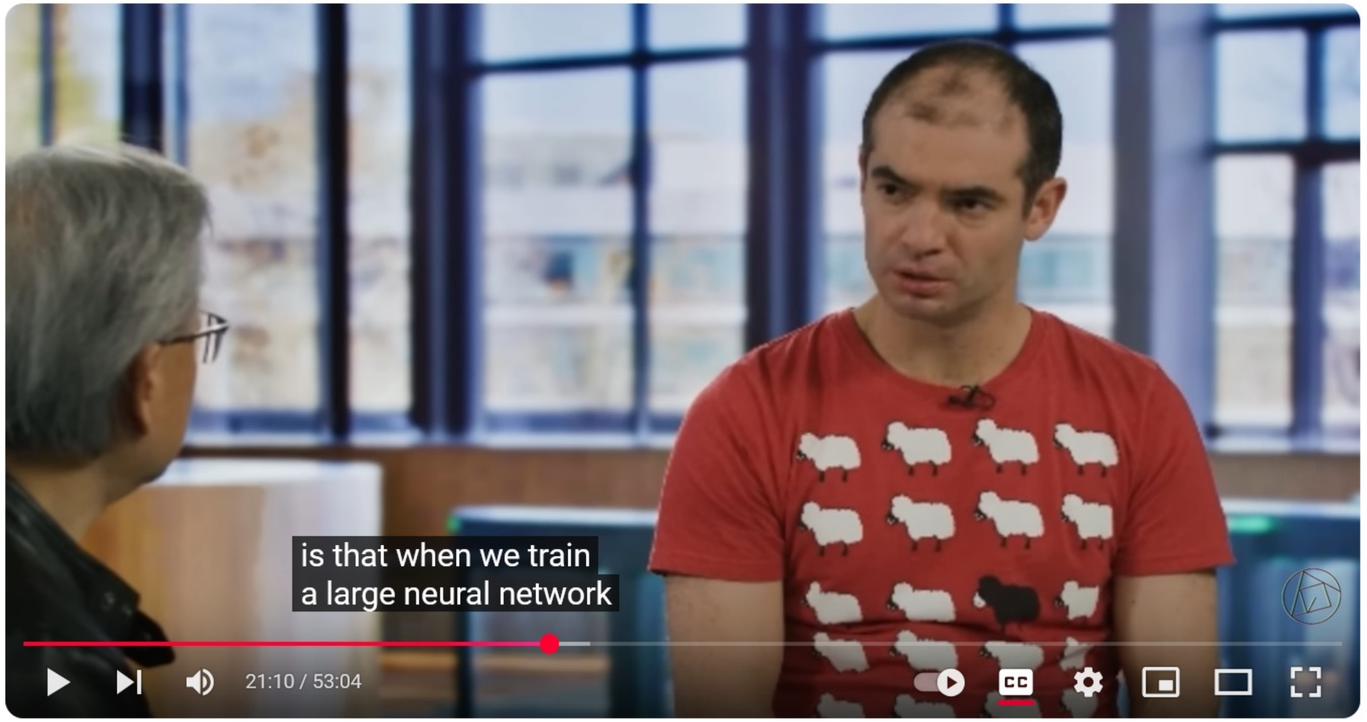
Timnit Gebru*
timnit@blackinai.org
Black in AI
Palo Alto, CA, USA

Angelina McMillan-Major
aymm@uw.edu
University of Washington
Seattle, WA, USA

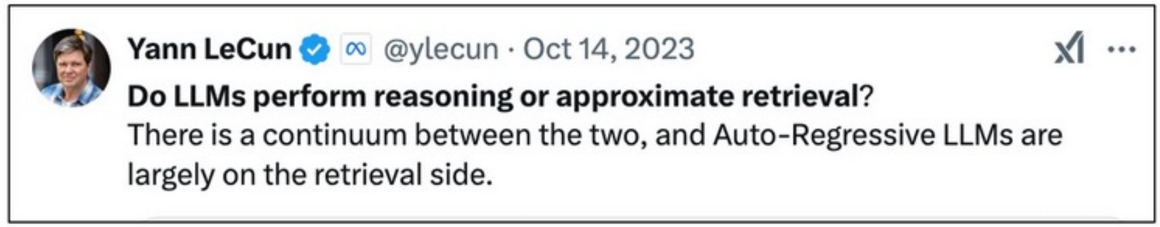
Shmargaret Shmitchell
shmargaret.shmitchell@gmail.com
The Aether

https://dl.acm.org/doi/epdf/10.1145/3442188.3445922

There is a debate in the field on whether LLMs reason (use a “world model”).



CONFERENCE JENSEN HUANG (NVIDIA) and ILYA SUTSKEVER (OPEN AI).AI Conferenza SUB ITA



“When we train a large neural network to accurately predict the next word in lots of different texts...it is learning a world model.... This text is actually a projection of the world.... What the neural network is learning is more and more aspects of the world, of people, of the human conditions, their hopes, dreams, and motivations...the neural network learns a compressed, abstract, usable representation of that.”

Code Generation: Threat or Menace?

Instruction:

Write code to create a list of all prime numbers between 2 and 100.

Input:

N/A

Output:

```
def find_primes(n):
    prime_list = [2]
    for number in range(2, n + 1):
        is_prime = True
        for k in range(2, number):
            if number % k == 0:
                is_prime = False
        if is_prime:
            prime_list.append(number)
    return prime_list
```

Instruction:

Generate a snippet of code to extract all the URLs from the given string.

Input:

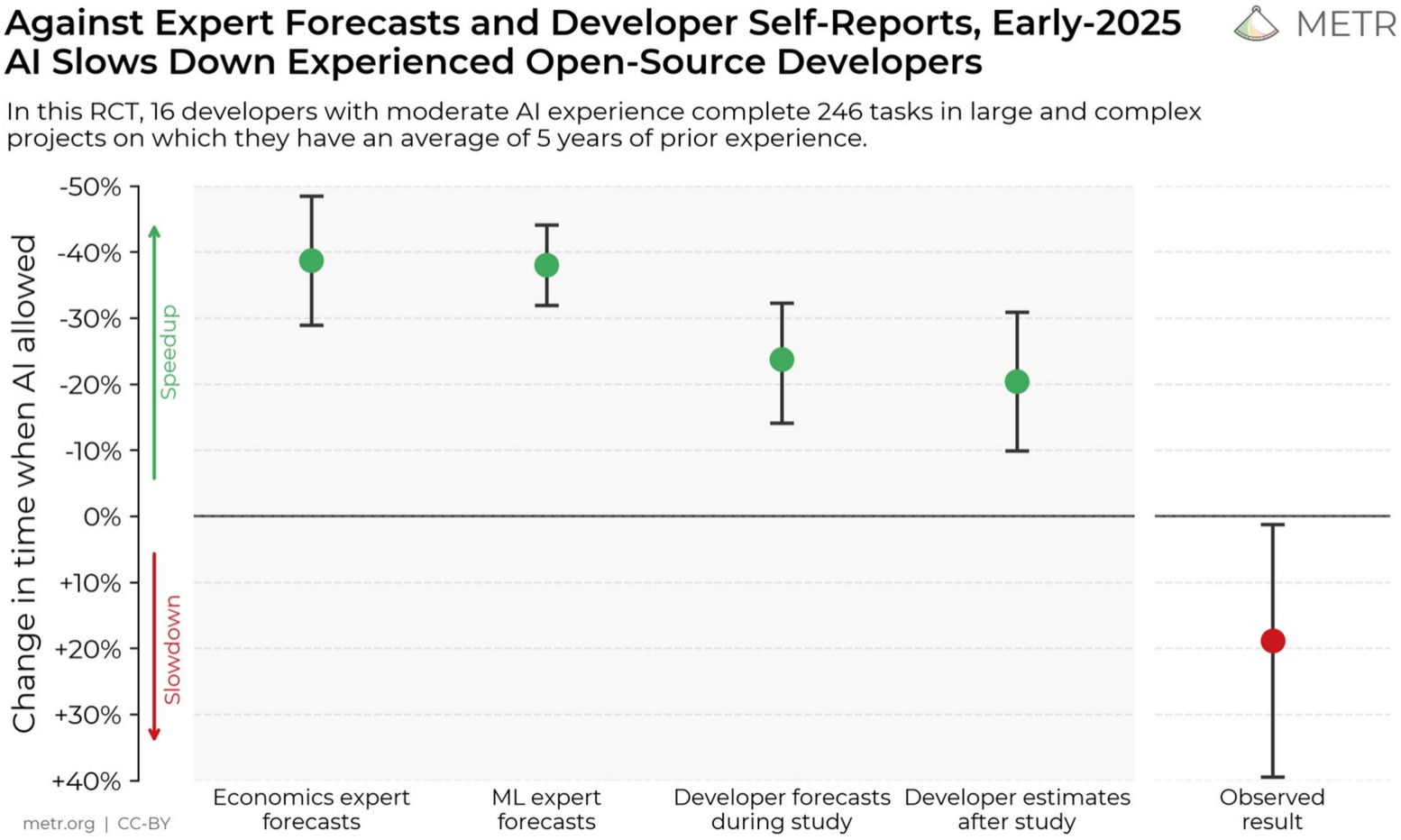
This string contains some urls such as https://www.google.com and https://www.facebook.com.

Output:

```
import re

string = "This string contains some
urls such as https://www.google.com and
https://www.facebook.com."

urls = re.findall('http[s]?://(?:[a-zA-
Z]|[0-9]|[$-_@.&+]|[!*\(\),]|(?:%[0-9a-
fA-F][0-9a-fA-F]))+', string)

print(urls)
```

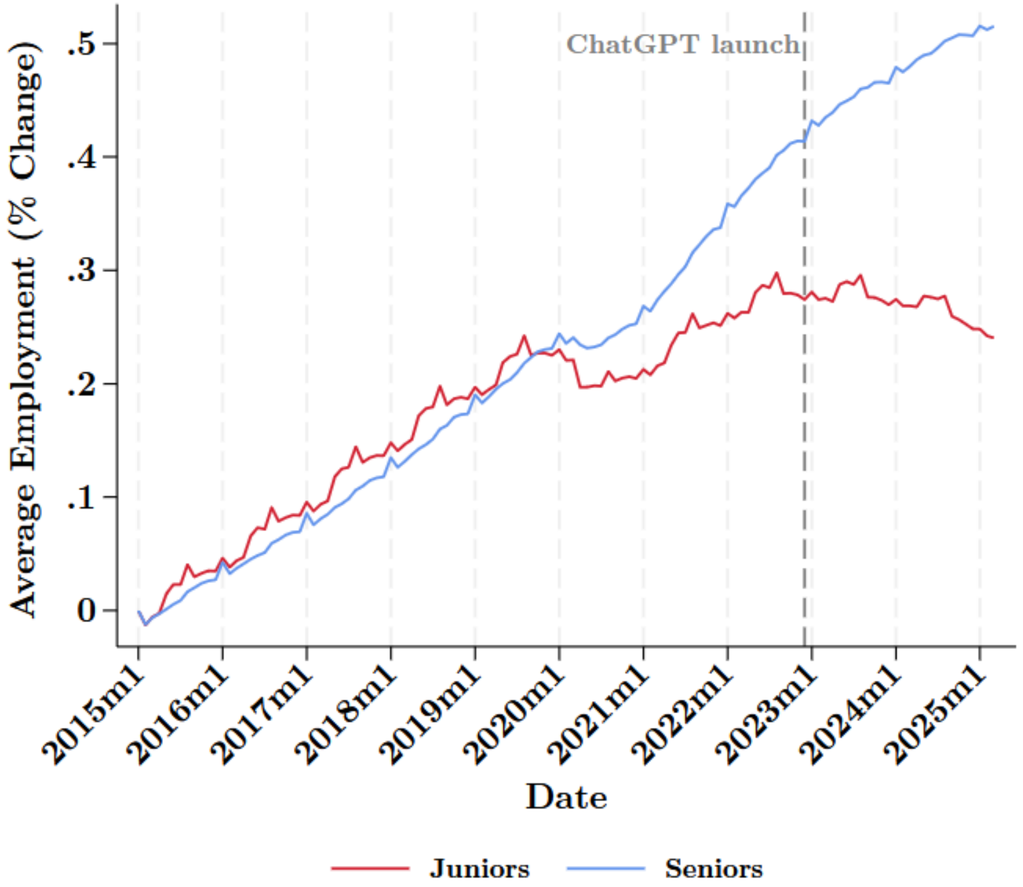
Juyong Jiang, Fan Wang, Jiasi Shen, Sungju Kim, and Sunghun Kim. 2026. A Survey on Large Language Models for Code Generation. ACM Trans. Softw. Eng. Methodol. 35, 2, Article 58 (February 2026), 72 pages. https://doi.org/10.1145/3747588

METR measured the impact on productivity on experienced developers (2025).



https://metr.org/blog/2025-07-10-early-2025-ai-experienced-os-dev-study/

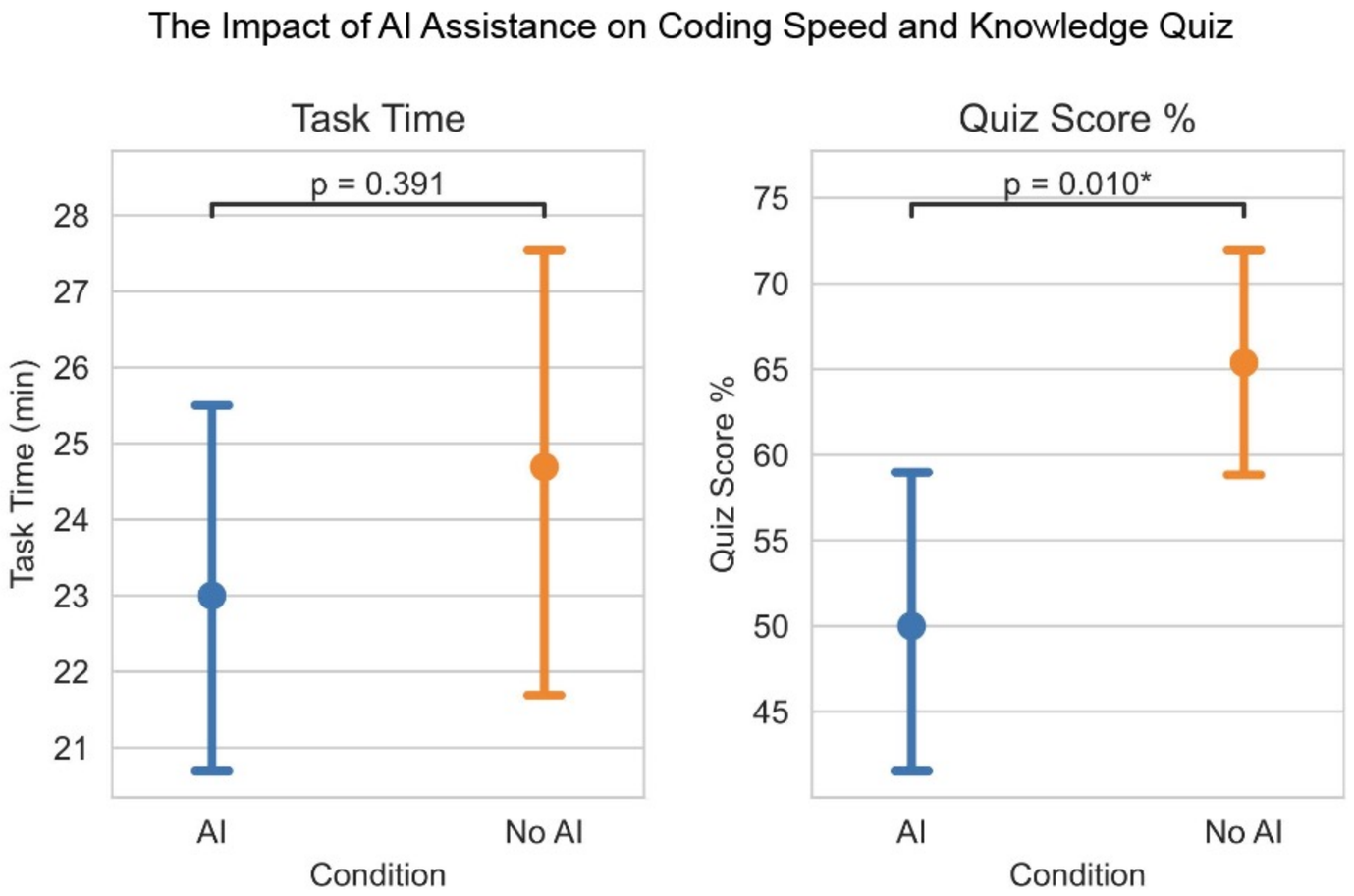
Generative AI may be affecting junior-level hiring (Sept. 2025)



"Generative AI as Seniority-Biased Technological Change: Evidence from U.S. Résumé and Job Posting Data," Seyed Mahdi Hosseini Maasoum, Harvard University, Guy Lichtinger, Harvard University Department of Economics. 8 Sep 2025.

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5425555

Anthropic found losses in comprehension in AI users (January 2026)



https://www.anthropic.com/research/AI-assistance-coding-skills

https://arxiv.org/pdf/2601.20245

Is AI a “Normal” Technology?

AI as Normal Technology

An alternative to the vision of AI as a potential superintelligence

By Arvind Narayanan and Sayash Kapoor

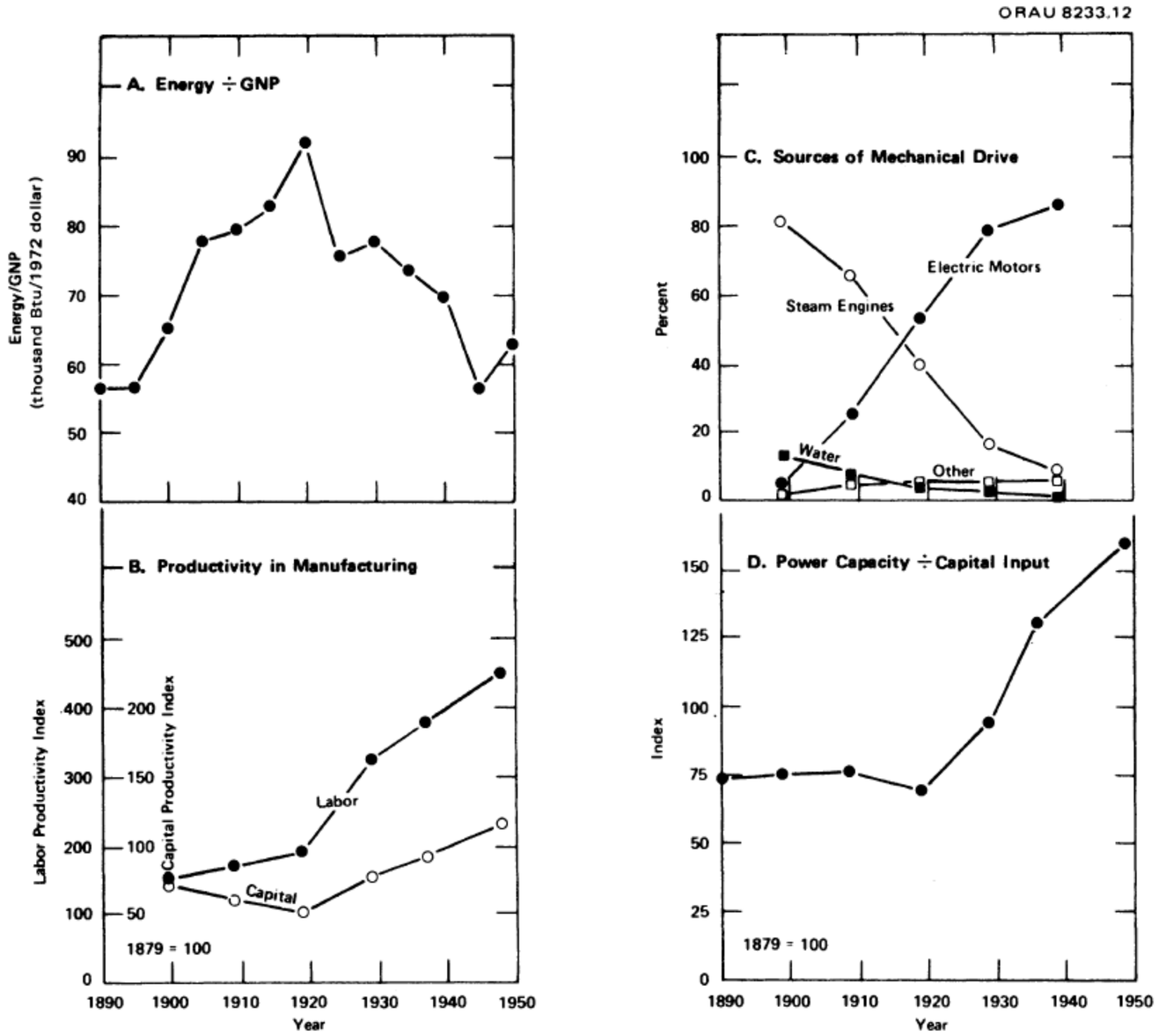
April 15, 2025

The statement “AI is normal technology” is three things: a *description* of current AI, a *prediction* about the foreseeable future of AI, and a *prescription* about how we should treat it. We view AI as a tool that we can and should remain in control of, and we argue that this goal does not require drastic policy interventions or technical breakthroughs. We do not think that viewing AI as a humanlike intelligence is currently accurate or useful for understanding its societal impacts, nor is it likely to be in our vision of the future.[2]

https://knightcolumbia.org/content/ai-as-normal-technology

Factories before electrification were driven off a single shaft powered by wind or water

A common comparison is that AI is the “new electricity.”



Next week...

When requesting help

If you have a question about a grade or will have an absence:

- Please use Canvas for requesting help about grades
- Please be sure to create a new message thread if you are emailing about a new topic.
- Do not just reply to the previous message.

If you have a technical question about your lab:

- Please post to Ed Board.
- Please post both your **code that produced the problem** and the **error message that you saw.**
- If you only post one, we will need you to ask you to post the other.

If you have posted a problem:

- Please leave your Instance running so we can log into it.
- If you have “access on,” we will assume we can log into your instance to answer your question.

Next Week's Reading

GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models

Iman Mirzadeh† Keivan Alizadeh Hooman Shahrokhi*
Oncel Tuzel Samy Bengio Mehrdad Farajtabar†

Apple

Abstract

Recent advancements in Large Language Models (LLMs) have sparked interest in their formal reasoning capabilities, particularly in mathematics. The GSM8K benchmark is widely used to assess the mathematical reasoning of models on grade-school-level questions. While the performance of LLMs on GSM8K has significantly improved in recent years, it remains unclear whether their mathematical reasoning capabilities have genuinely advanced, raising questions about the reliability of the reported metrics. To address these concerns, we conduct a large-scale study on several state-of-the-art open and closed models. To overcome the limitations of existing evaluations, we introduce GSM-Symbolic, an improved benchmark created from symbolic templates that allow for the generation of a diverse set of questions. GSM-Symbolic enables more controllable evaluations, providing key insights and more reliable metrics for measuring the reasoning capabilities of models.Our findings reveal that LLMs exhibit noticeable variance when responding to different instantiations of the same question. Specifically, the performance of all models declines when only the numerical values in the question are altered in the GSM-Symbolic benchmark. Furthermore, we investigate the fragility of mathematical reasoning in these models and demonstrate that their performance significantly deteriorates as the number of clauses in a question increases. We hypothesize that this decline is due to the fact that current LLMs are not capable of genuine logical reasoning; instead, they attempt to replicate the reasoning steps observed in their training data. When we add a single clause that appears relevant to the question, we observe significant performance drops (up to 65%) across all state-of-the-art models, even though the added clause does not contribute to the reasoning chain needed to reach the final answer. Overall, our work provides a more nuanced understanding of LLMs' capabilities and limitations in mathematical reasoning.

Marcus on AI

LLMs don't do formal reasoning - and that is a HUGE problem

Important new study from Apple

GARY MARCUS
OCT 11, 2024

A superb new article on LLMs from six AI researchers at Apple who were brave enough to challenge the dominant paradigm has just come out.

Everyone actively working with AI should read it, or at least this terrific X thread by senior author, Mehrdad Farajtabar, that summarizes what they observed. One key passage:

> "we found no evidence of formal reasoning in language models .... Their behavior is better explained by sophisticated pattern matching—so fragile, in fact, that changing names can alter results by ~10%!"

One particularly damning result was a new task the Apple team developed, called GSM-NoOp

For next week:

- Read Mirzadeh
  - Comment on Perusall
  - Take the weekly reading quiz
- Participate in Ed Board
- Finish Lab 6
- Finish your taxes (US people)