

Generative AI and Large Language Models: How We Got Here

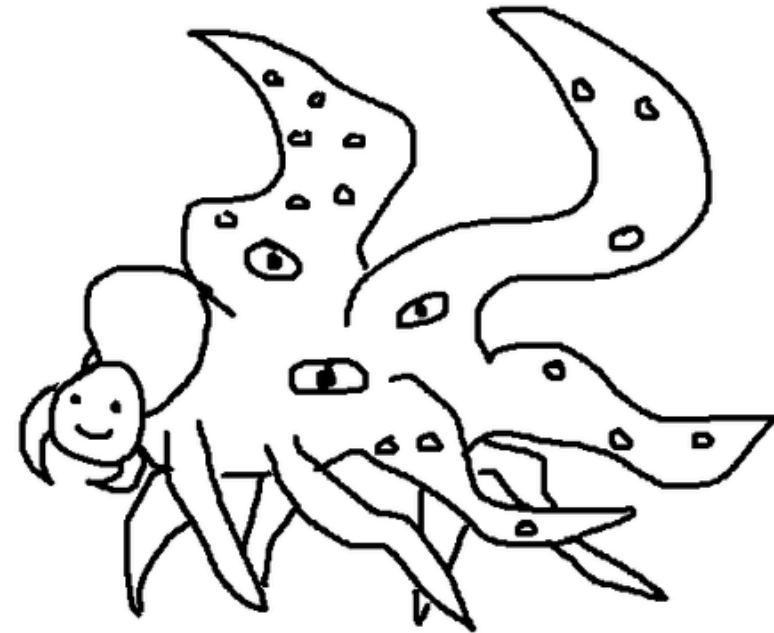
Michael White
Department of Linguistics
The Ohio State University

ASC Teaching Forum —
Guiding Principles and Innovative Use Case for Teaching with A.I.
March 18, 2024

So how does ChatGPT actually work?

- The short answer: we don't really know!
- The longer answer: we do actually know how it's designed and how it's trained, but as a massive **black box**, we understand little of what it's really doing
- Analogy: I could tell you the pixel values of every frame of the movie *Barbie*, but that would hardly tell you whether it's worth seeing!

GPT-3 + RLHF



Shoggoth meme ([NYT](#))

A brief history of ChatGPT and friends

- Large language models (LLMs) are enormous neural network models trained on massive datasets to predict words sequences
- **Instruction-tuned** models like ChatGPT are further trained to answer questions and follow instructions (RLHF)
- Neural networks go back to the 80s and 90s, but did not work particularly well at the time
- The nascent field of **machine learning** instead focused on mathematically better understood models (90s–00s)
- Work in AI was turned on its head around 2010 when more data and more compute led “deep” neural network models to achieve unexpected performance breakthroughs — first in vision, then in language

Word embeddings combat data sparsity

Distributional information about a word can be used to derive a numeric, vector-based representation of its meaning: a **word embedding**

“You shall know a word by the company it keeps” (Firth, 1957)

Representing words as vectors means we can think of them as points in a multidimensional semantic space

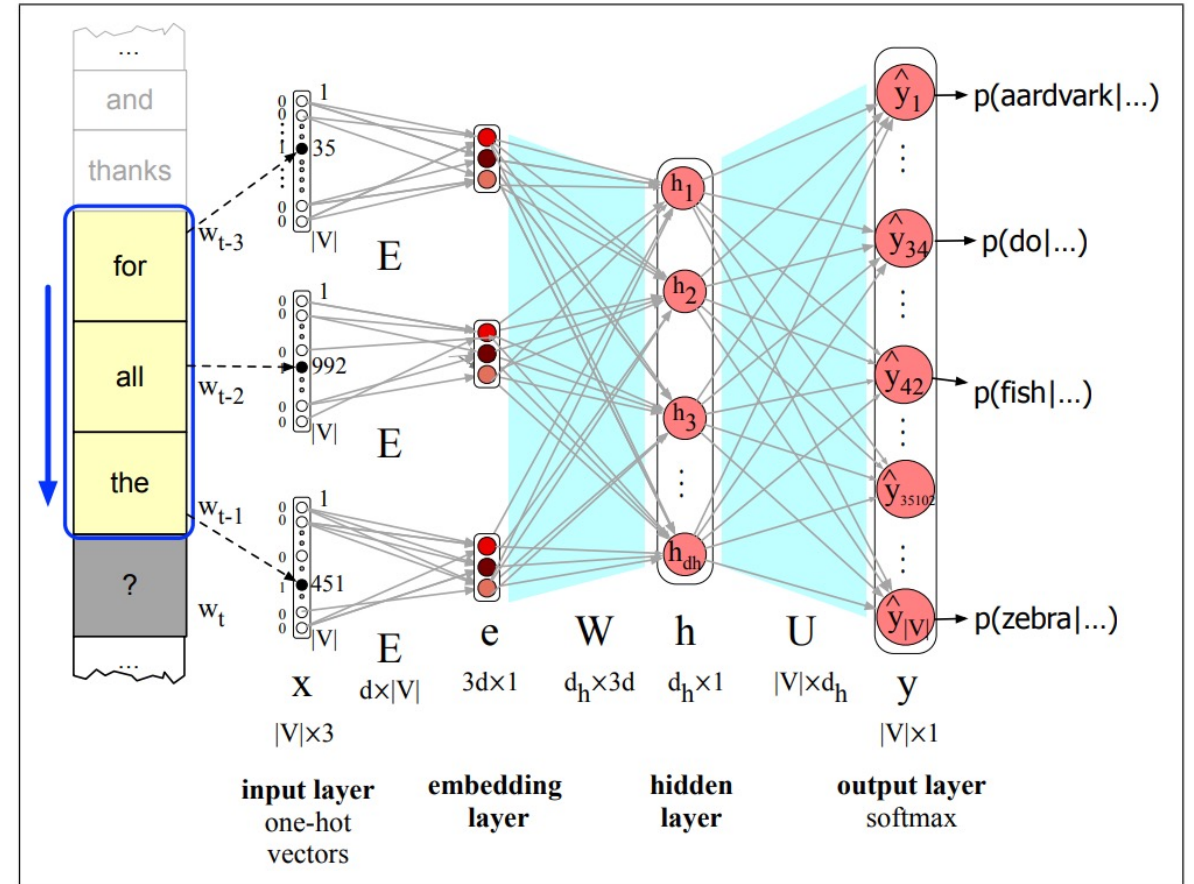
Nearby words have similar meanings

Example two-dimensional representations:



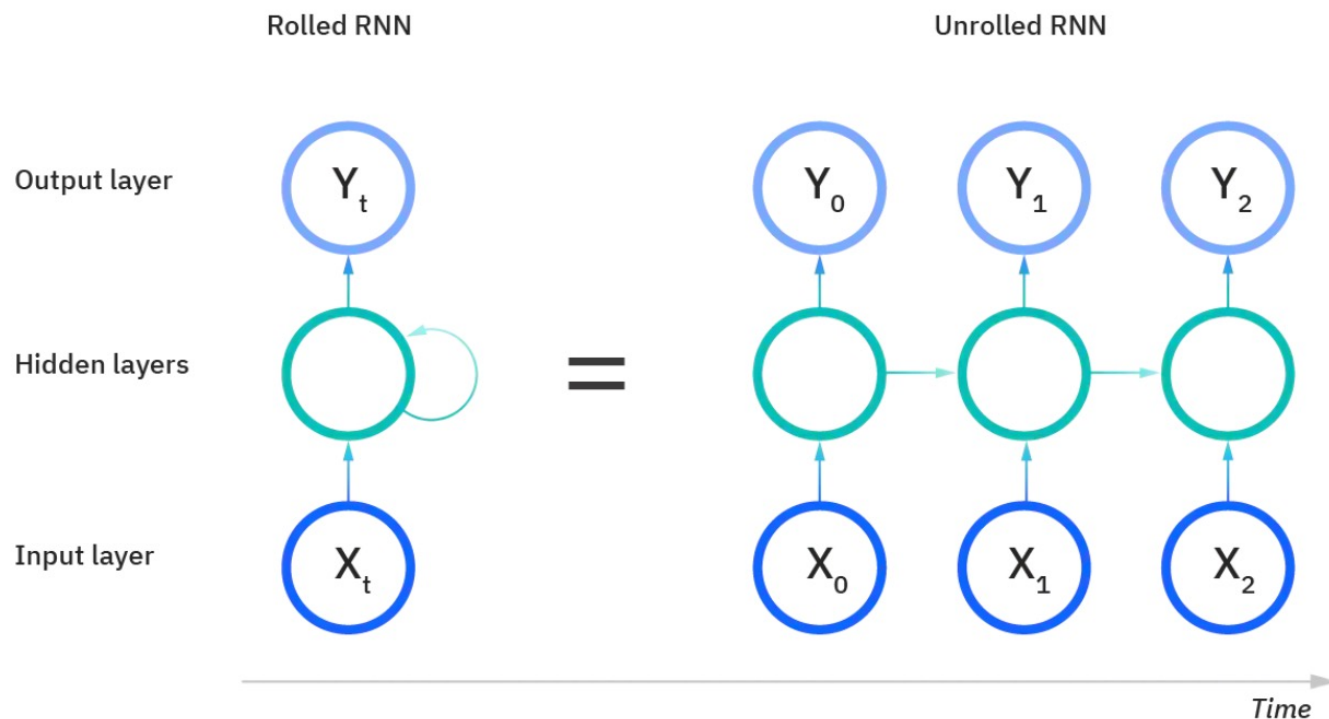
Feedforward networks for language modeling

- Feedforward networks can be used for language modeling
- Input: **embeddings** for the previous N words
 - N is some fixed value (e.g. $N=3$ in the figure)
- Output: probability distribution for the next word in the sequence



Recurrent neural networks (Elman 1990)

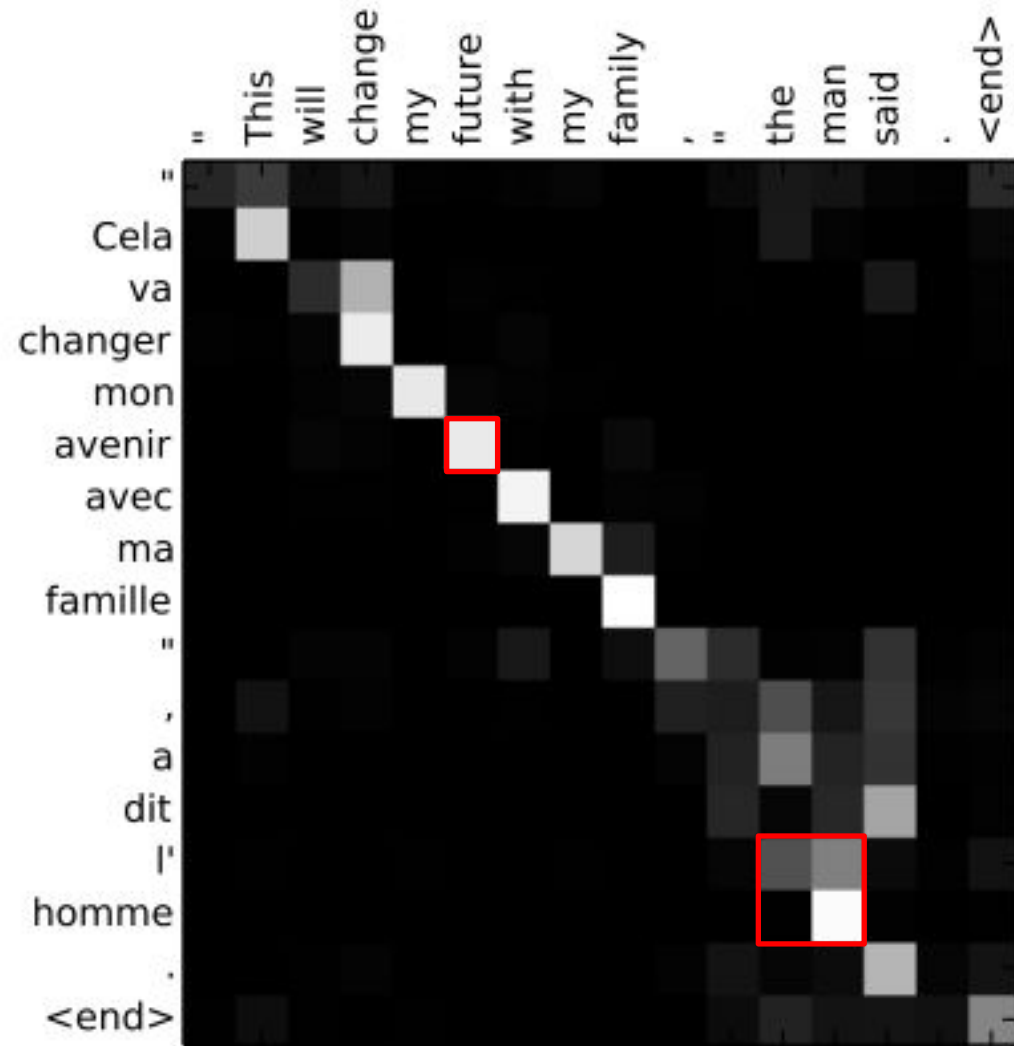
- Can condition the next word in a sequence on an unbounded amount of context



<https://www.ibm.com/cloud/learn/recurrent-neural-networks>

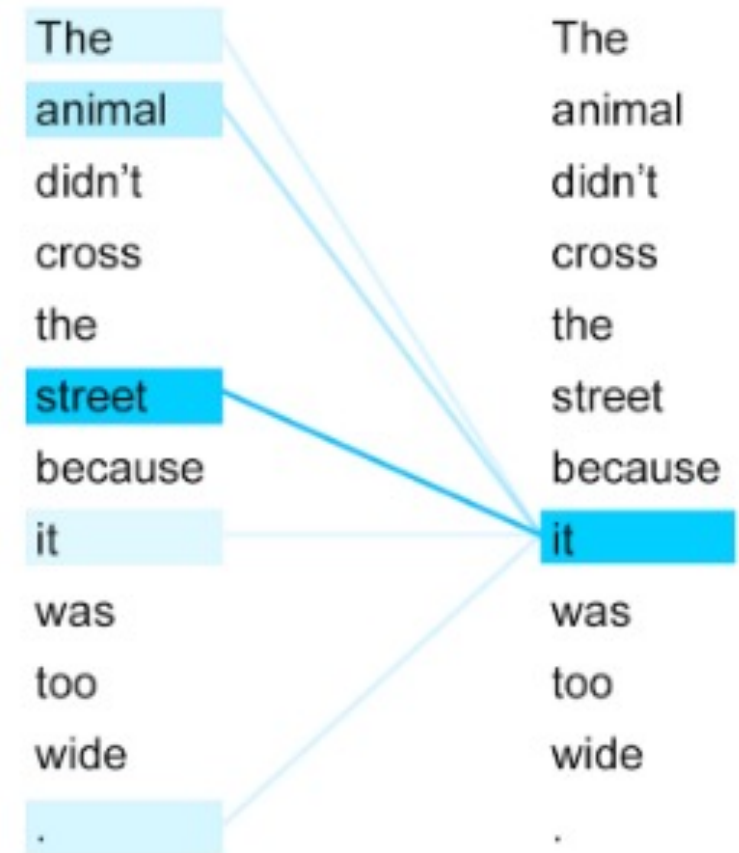
Attention and Alignment

- The **attention** scores play a similar role to alignment in earlier statistical models
- Bahdanau, Cho and Bengio (2014), “Neural machine translation by jointly learning to align and translate”
- Most words are aligned one-to-one, but some are many-to-many

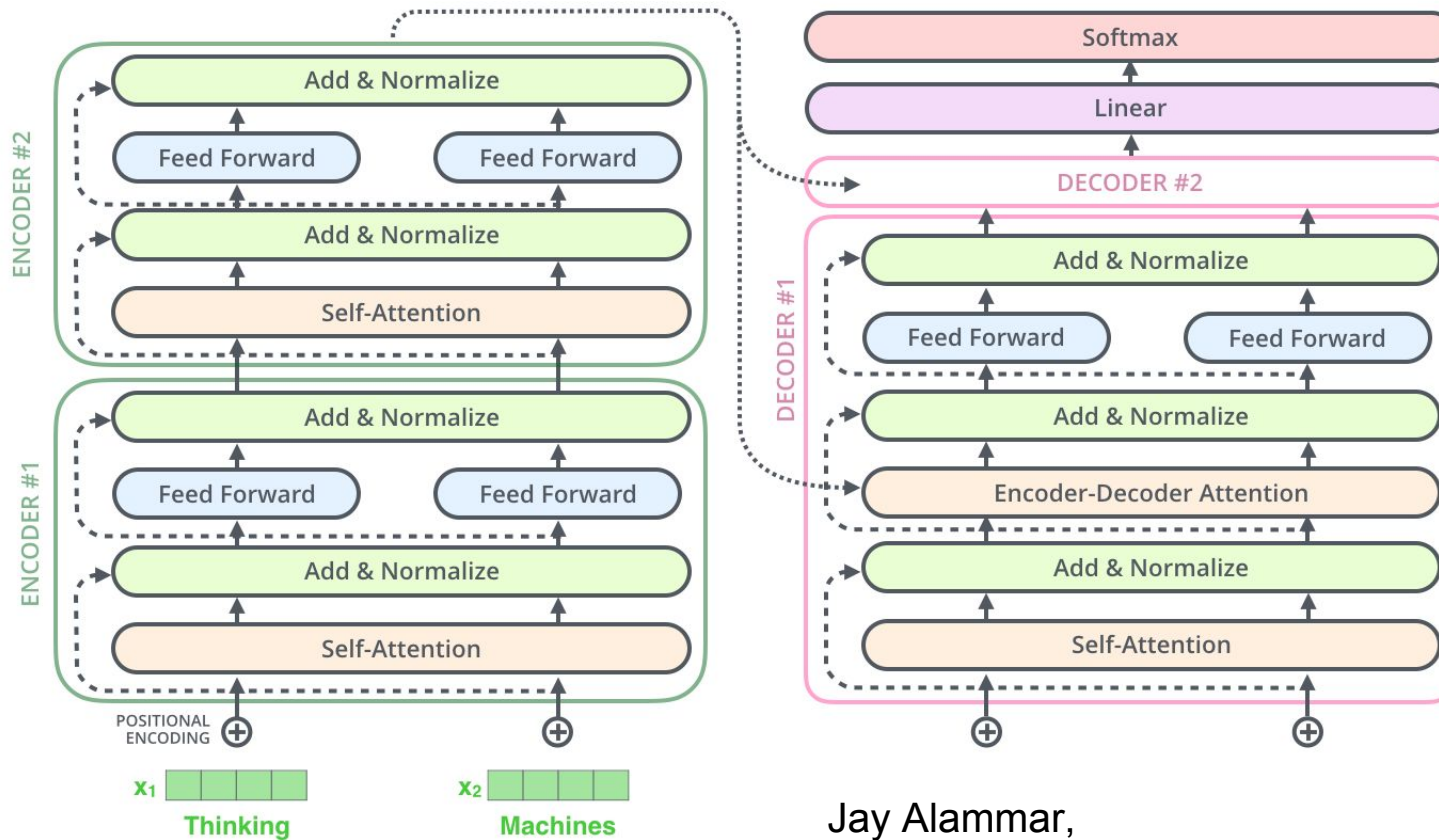


“Attention is all you need” (Vaswani et al., 2017)

- The **Transformer** architecture for neural network models was introduced in 2017
- Like recurrent neural networks, Transformers are designed for sequential data
- Gets rid of RNN, just keeping attention mechanism — “Attention is all you need”
- This enables data parallelism in training, making it possible to use GPUs to train on massive data

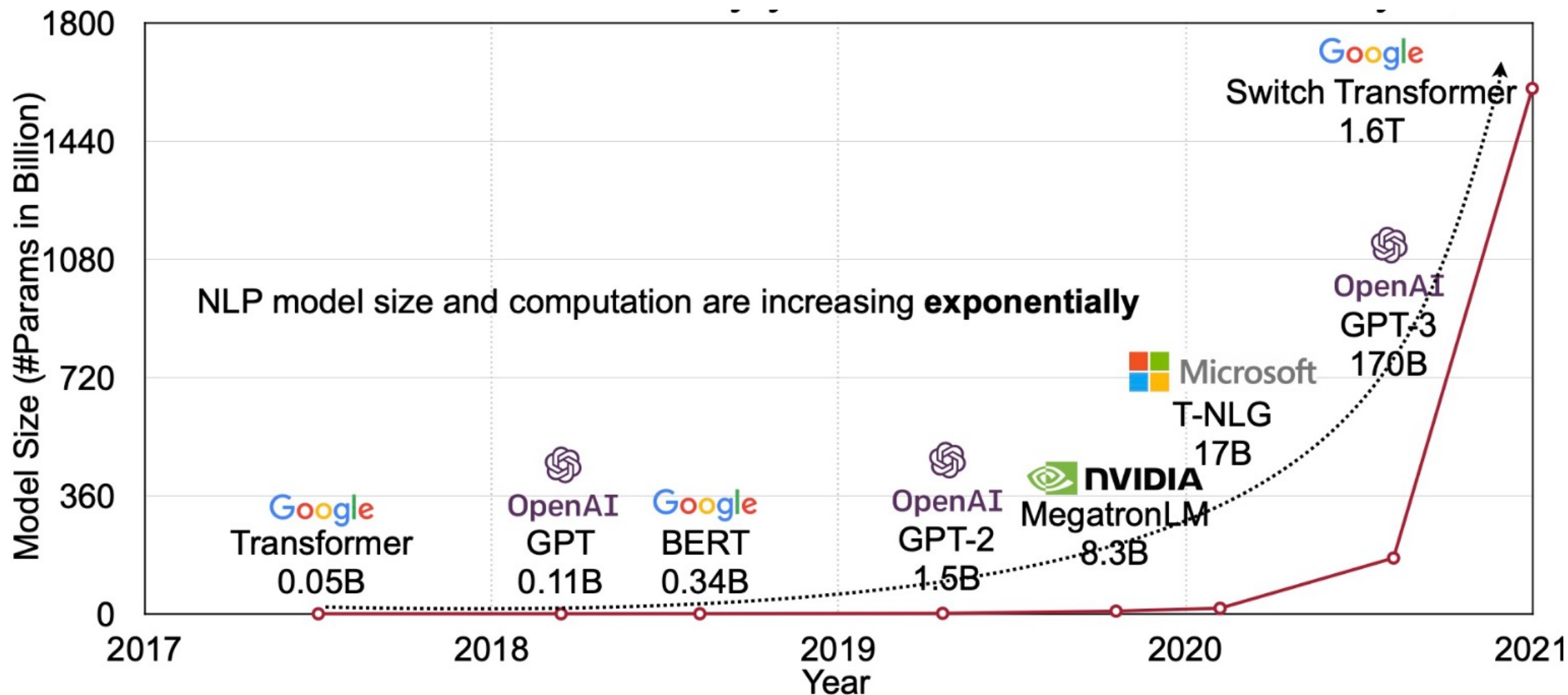


Transformer Architecture



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

“Scale is all you need”



LLMs have emergent capabilities

- More recent large language models (LLMs) such as GPT-3 and GPT-4, used in OpenAI's ChatGPT, simplify the architecture to just a **decoder pre-trained on next-word prediction**
- Instruction-tuning from human feedback (RLHF) also added
- Greater scale has enabled **emergent capabilities** such as the ability to do **in-context learning**
 - That is, these models can carry out tasks just given instructions and a few examples (a.k.a. **few-shot learning**), with no supervised fine-tuning
 - Or just given the instructions themselves! (a.k.a. **zero-shot learning**)
- These models generate responses just via LM sampling — but what's represented in their hidden states **remains largely a mystery**

An emergent
ability:
asterisk
corrections

DO YOU HAVE ANY
WEEKEND PLANS?

I'M GONNA RIDE A HORSE
ON THE BEACH AT DAWN

*EAT

*3AM

*COUCH

*PIZZA

I LIKE HOW WE CAN DO CORRECTIONS
IN TEXT CHAT BY APPENDING WORDS
WITH ASTERISKS AND OUR BRAINS
JUST FIGURE OUT WHERE THEY GO.

ChatGPT asterisk corrections

Me

Consider the following message exchange:

A: Do you have any weekend plans?

B: I'm going to ride a horse on the beach at dawn.

...

Given these asterisk corrections, what was B's actual reply?


ChatGPT

Based on the corrections provided by B, their actual reply would be:

B: I'm going to eat pizza on the couch at 3 AM.

But all is not well ...

Terminator scenario: $P(\text{doom}) > 0$

Bender et al.'s (2021) *Stochastic Parrots*  [paper](#) argues that LLMs do not actually understand language and many potential harms follow (though doom scenario is a distraction)

