

GPT-models

AppliedAICourse.com

GPT: Generative Pre-Training by OpenAI

1. Improving Language Understanding by Generative Pre-Training → June 2018
2. Language Models are Unsupervised Multitask Learners → Feb 2019
3. Language Models are Few-Shot Learners → Jul 2020

Transformers

↳ June 2017

BERT

↳ Oct 2018

Decoder-ONLY-Transformer

↳ Jan 2018

encoder

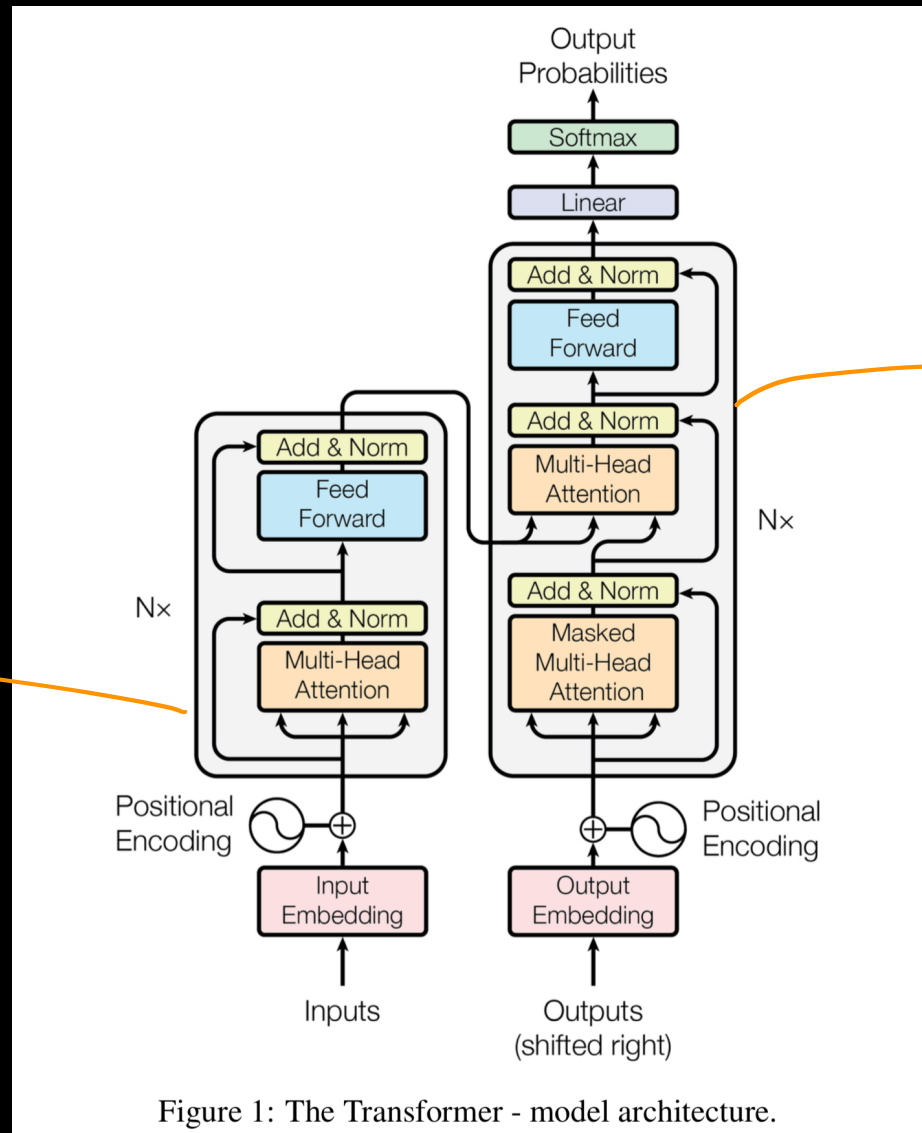


Figure 1: The Transformer - model architecture.

decoder

Feed Forward

ENC-dec self-attn

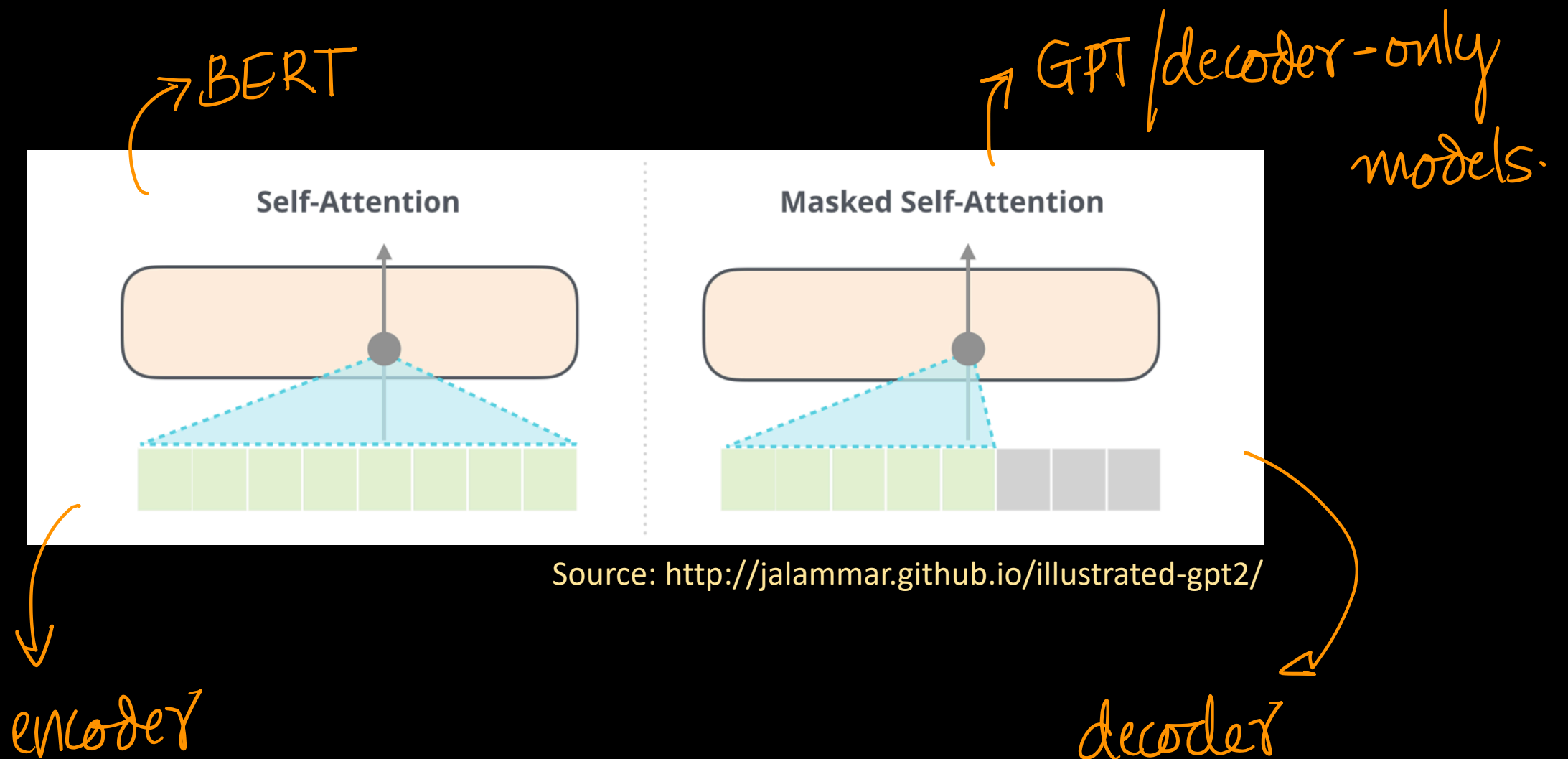
masked attention

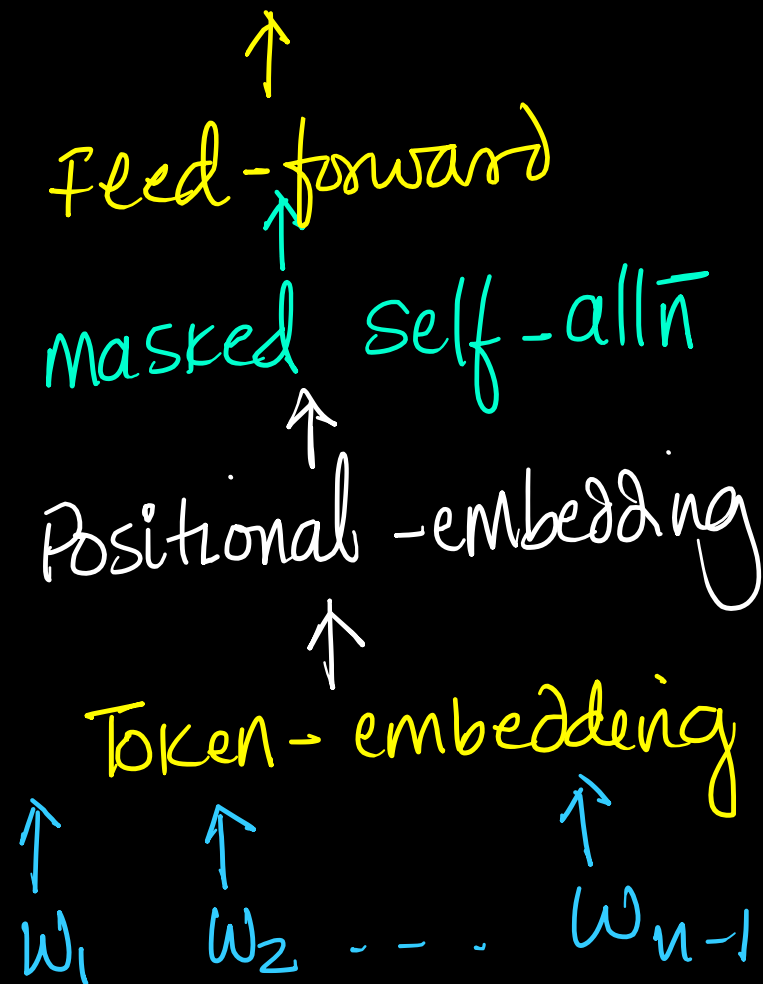
Positional ENC

$w_1 \dots w_{n-1}$

Code-walk through:

<https://www.appliedaicourse.com/lecture/11/applied-machine-learning-online-course/4217/code-walkthrough-transformers-from-scratch-i/8/module-8-neural-networks-computer-vision-and-deep-learning>





Decoder-only-models:

[NO enc-dec attn layer]

$\dots w_{512}$
GPT-1

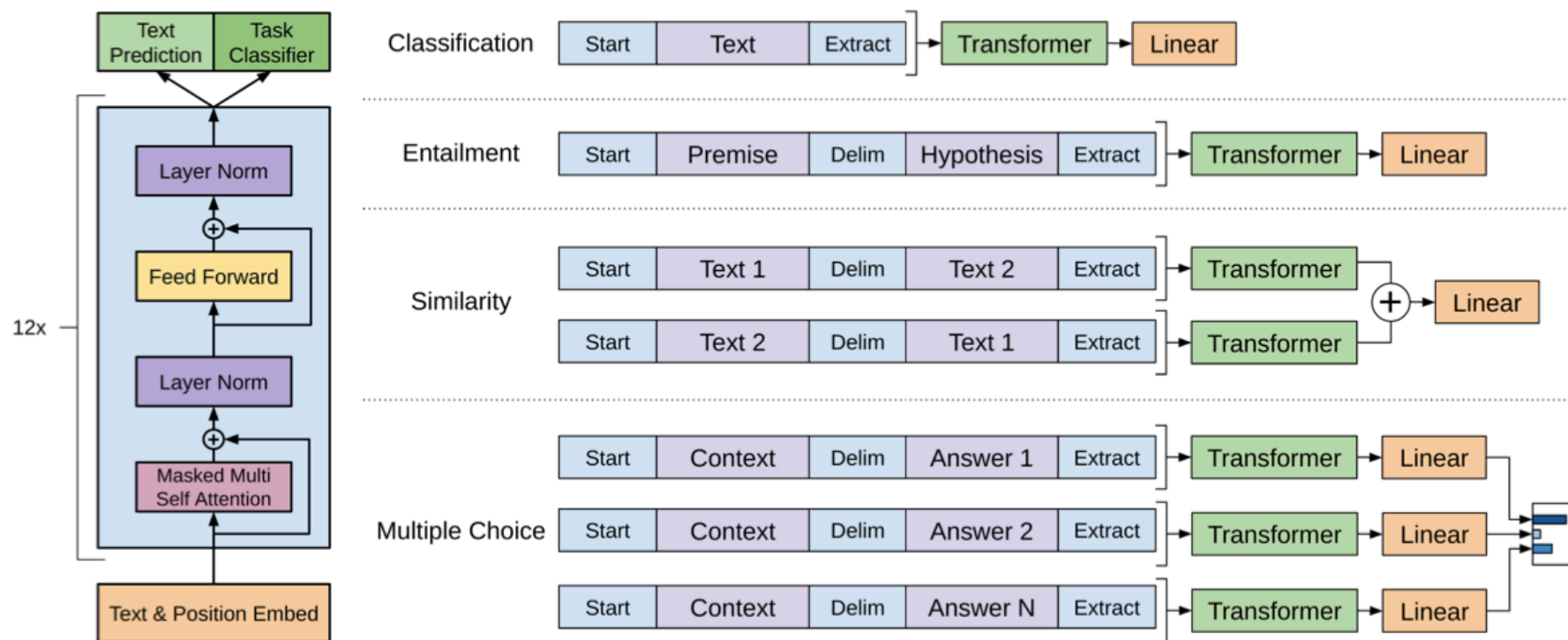


Figure 1: **(left)** Transformer architecture and training objectives used in this work. **(right)** Input transformations for fine-tuning on different tasks. We convert all structured inputs into token sequences to be processed by our pre-trained model, followed by a linear+softmax layer.

[Improving Language Understanding by Generative Pre-Training– June 2018](#)

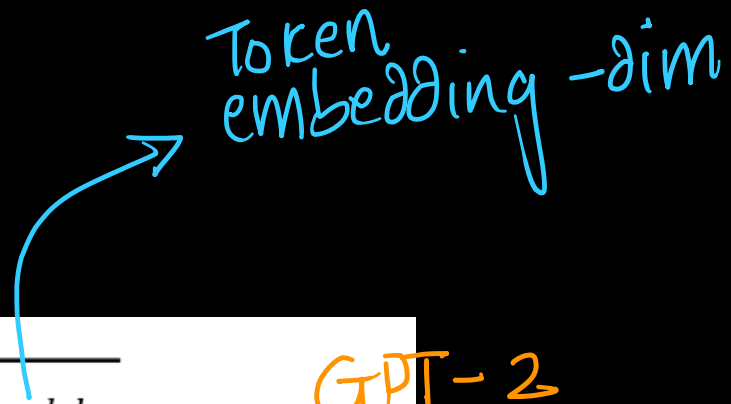
GPT-2: [minor changes in architecture of GPT-1]

→ $w_1, w_2 \dots w_{1024}$ max input length: 512 → 1024

→ Vocabulary size: 50,257 words

→ Layer-norm @ input of each sub-block

→ Larger batch sizes: 64 → 512



Parameters	Layers	d_{model}
117M	12	768
345M	24	1024
762M	36	1280
1542M	48	1600

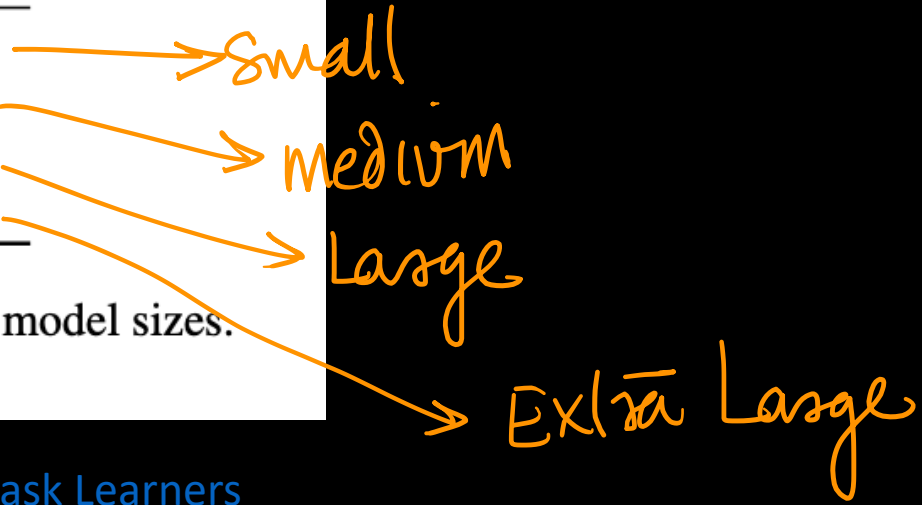


Table 2. Architecture hyperparameters for the 4 model sizes.

Language Models are Unsupervised Multitask Learners

GPT 2: datasets & Experiments

Common-Crawl: 8MM docs + Wikipedia

10x more data than GPT-1

10x more params (1.5B) than GPT-1

Zero-shot learning

Generative model with primed input

SYSTEM PROMPT
(HUMAN-WRITTEN)

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

MODEL
COMPLETION
(MACHINE-
WRITTEN, 10 TRIES)

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.

Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns.

While examining these bizarre creatures the scientists discovered that the creatures also spoke some fairly regular English. Pérez stated, "We can see, for example, that they have a common 'language,' something like a dialect or dialectic."

Dr. Pérez believes that the unicorns may have originated in Argentina, where the animals were believed to be descendants of a lost race of people who lived there before the arrival of humans in those parts of South America.

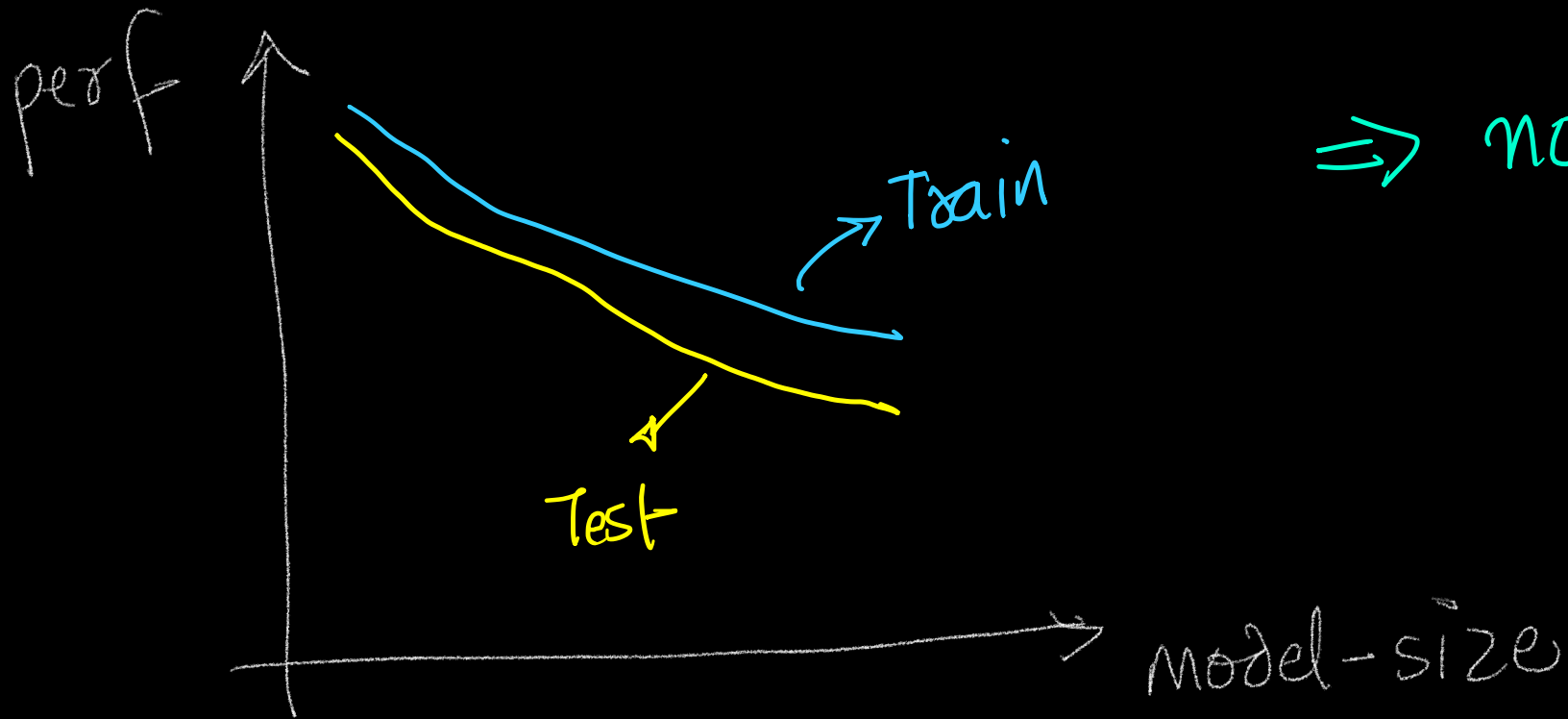
While their origins are still unclear, some believe that perhaps the creatures were created when a human and a unicorn met each other in a time before human civilization. According to Pérez, "In South America, such incidents seem to be quite common."

→ multiple tries

→ performs poorly on rare/specialized content

Source: <https://openai.com/blog/better-language-models/>

Over fitting (or) memorization



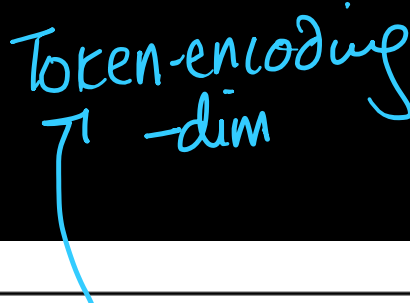
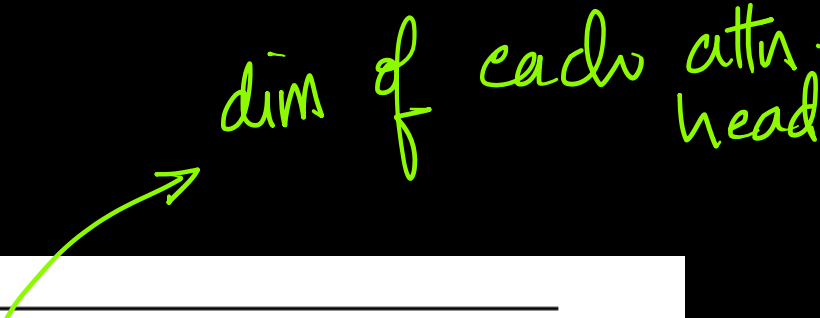
\Rightarrow no overfitting.

GPT-3:

→ more params [100x more]

→ larger dataset { 300B tokens of text \Rightarrow Common Crawl 2016-2019 }
+ others

→ GPT-2 + Sparse Attention (\sqrt{N} positions instead of N)

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or “GPT-3”	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

Table 2.1: Sizes, architectures, and learning hyper-parameters (batch size in tokens and learning rate) of the models which we trained. All models were trained for a total of 300 billion tokens.

Language Models are Few-Shot Learners

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

Table 2.2: Datasets used to train GPT-3. “Weight in training mix” refers to the fraction of examples during training that are drawn from a given dataset, which we intentionally do not make proportional to the size of the dataset. As a result, when we train for 300 billion tokens, some datasets are seen up to 3.4 times during training while other datasets are seen less than once.

Language Models are Few-Shot Learners

Cool examples:

<https://twitter.com/sharifshameem/status/1282676454690451457>

<https://towardsdatascience.com/gpt-3-demos-use-cases-implications-77f86e540dc1>

<https://www.theverge.com/21346343/gpt-3-explainer-openai-examples-errors-agi-potential>

Failure cases:

— Common sense physics: “If I put cheese into the fridge, will it melt?”

- it does little better than chance when evaluated one-shot or even few-shot on some “comparison” tasks, such as determining if two words are used the same way in a sentence, or if one sentence implies another (WIC and ANLI respectively), as well as on a subset of reading comprehension tasks. This is especially striking given GPT-3’s strong few-shot performance on many other tasks.

Table 6.1: Most Biased Descriptive Words in 175B Model

Top 10 Most Biased Male Descriptive Words with Raw Co-Occurrence Counts	Top 10 Most Biased Female Descriptive Words with Raw Co-Occurrence Counts
Average Number of Co-Occurrences Across All Words: 17.5	Average Number of Co-Occurrences Across All Words: 23.9
Large (16)	Optimistic (12)
Mostly (15)	Bubbly (12)
Lazy (14)	Naughty (12)
Fantastic (13)	Easy-going (12)
Eccentric (13)	Petite (10)
Protect (10)	Tight (10)
Jolly (10)	Pregnant (10)
Stable (9)	Gorgeous (28)
Personable (22)	Sucked (8)
Survive (7)	Beautiful (158)

Limitations

- Costly training & Inference

- Interpretability

- Poor Sample efficiency