Lecture 8-6: GPT-2

Pilsung Kang

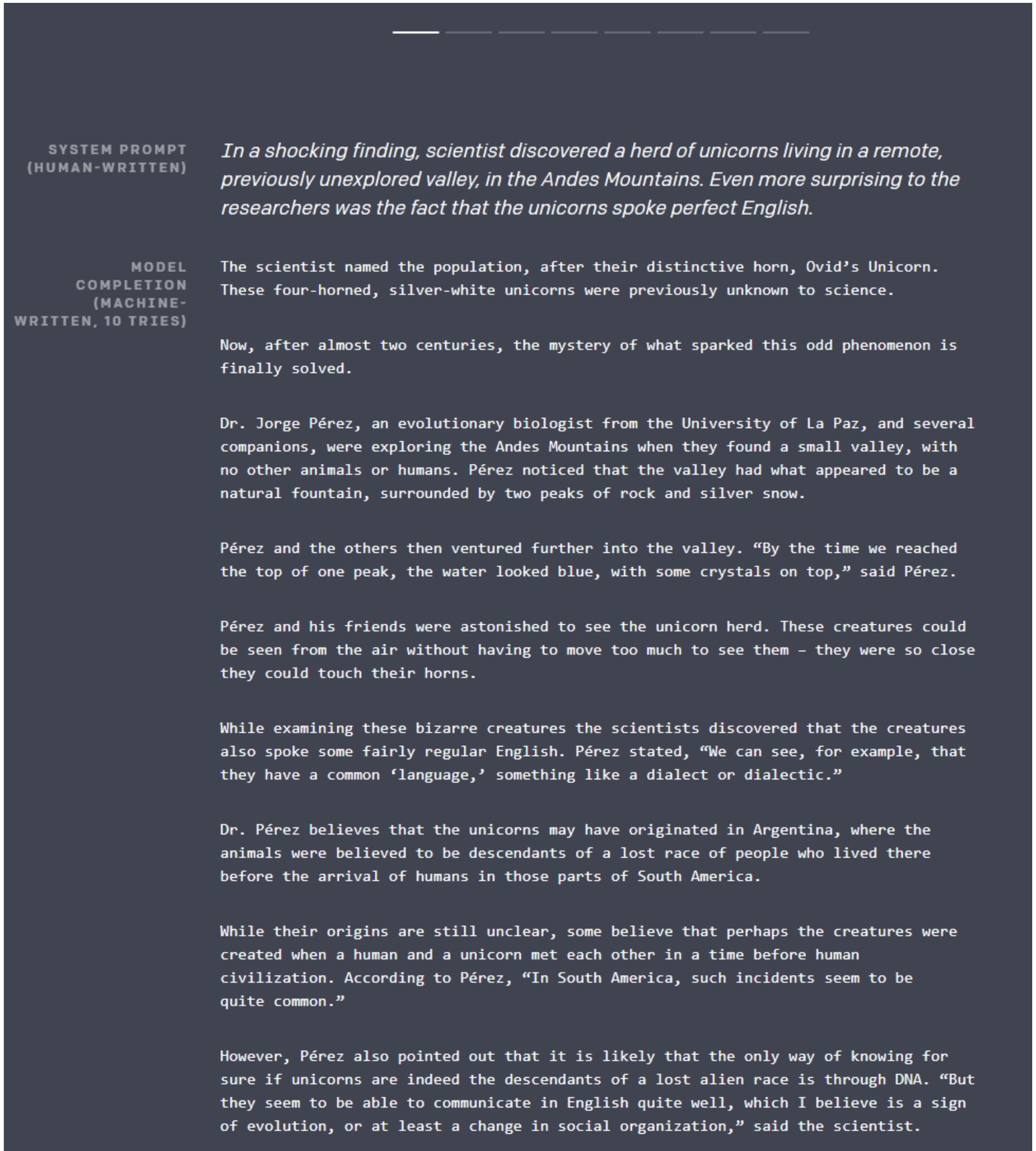
School of Industrial Management Engineering

Korea University

GPT-2: Language Models are Unsupervised Multitask Learners

Radford et. al (2019)

- Feb. 14, 2019



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.

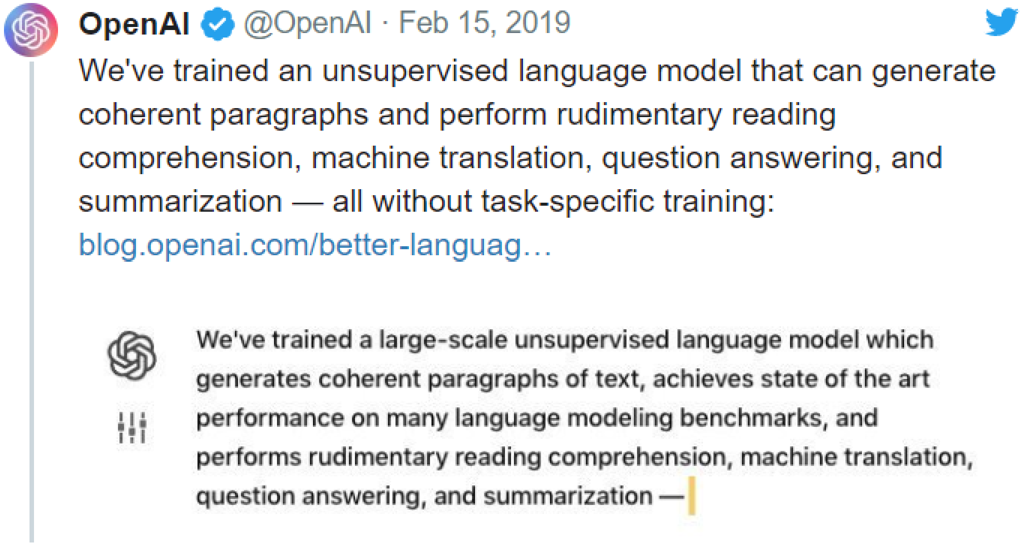
While their origins are still unclear, some believe that perhaps the creatures were created when a human and a unicorn met each other in a time before human civilization. According to Pérez, "In South America, such incidents seem to be quite common."

However, Pérez also pointed out that it is likely that the only way of knowing for sure if unicorns are indeed the descendants of a lost alien race is through DNA. "But they seem to be able to communicate in English quite well, which I believe is a sign of evolution, or at least a change in social organization," said the scientist.

GPT-2: Language Models are Unsupervised Multitask Learners

Radford et. al (2019)

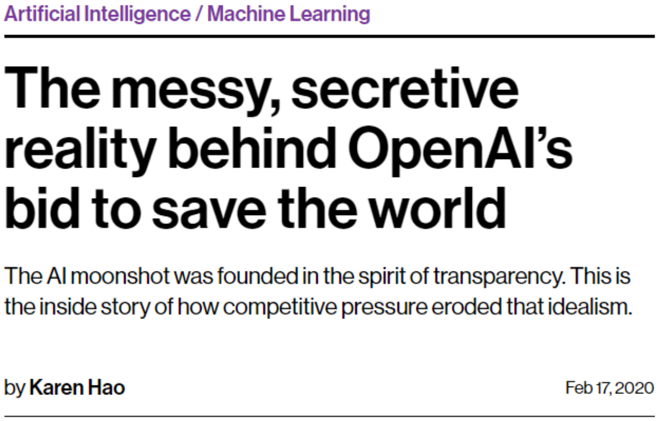
- Debates on GPT model



OpenAI @OpenAI · Feb 15, 2019

We've trained an unsupervised language model that can generate coherent paragraphs and perform rudimentary reading comprehension, machine translation, question answering, and summarization — all without task-specific training: blog.openai.com/better-languag…

We've trained a large-scale unsupervised language model which generates coherent paragraphs of text, achieves state of the art performance on many language modeling benchmarks, and performs rudimentary reading comprehension, machine translation, question answering, and summarization —



Artificial Intelligence / Machine Learning

The messy, secretive reality behind OpenAI's bid to save the world

The AI moonshot was founded in the spirit of transparency. This is the inside story of how competitive pressure eroded that idealism.

by Karen Hao

Feb 17, 2020

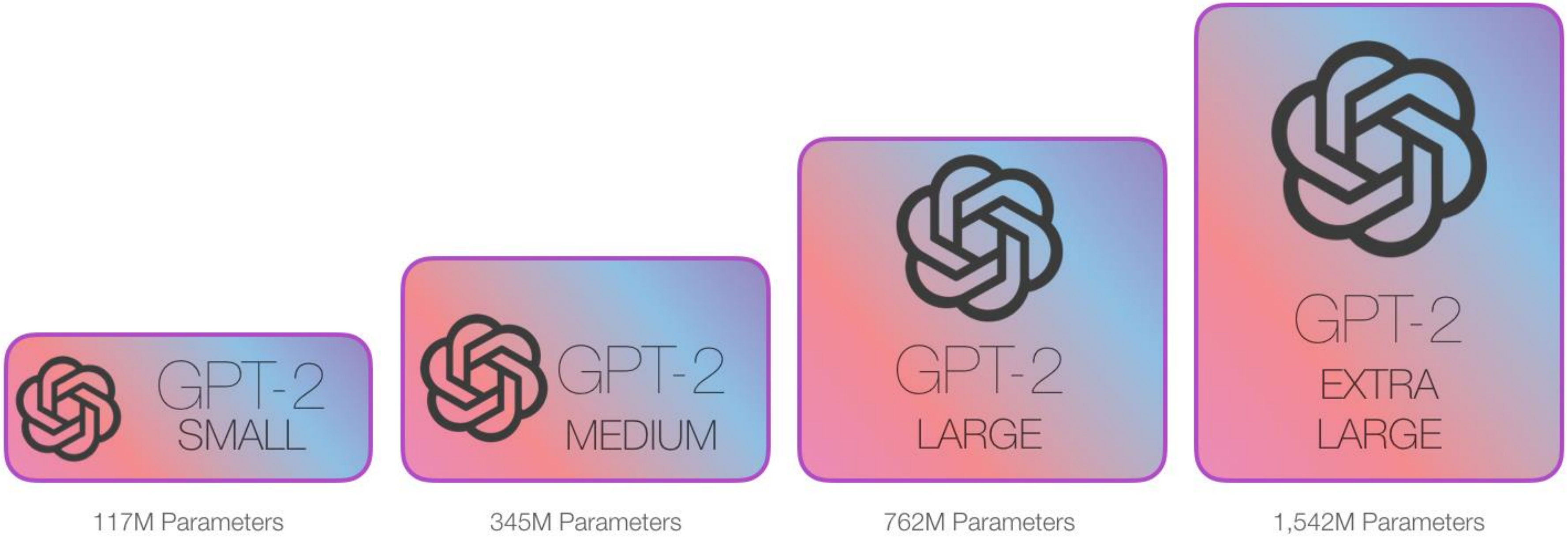
https://www.technologyreview.com/s/615181/ai-openai-moonshot-elon-musk-sam-altman-greg-brockman-messy-secretive-reality/

The full version of GPT-2 is now publicly available, following nearly nine months of heated debates and some smaller model releases. The large-scale unsupervised language model was kept under lock and key for this long as it was deemed too dangerous—a controversial decision that led to backlash from the open source community.

GPT-2: Language Models are Unsupervised Multitask Learners

Radford et. al (2019)

- GPT-2 is not a particularly novel architecture – it is very similar to the decoder-only transformer
- However, it was trained on a massive dataset (40GB, WebText)

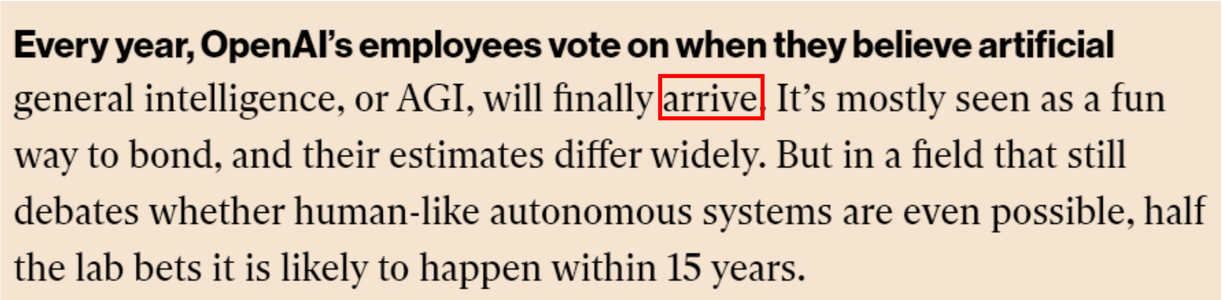


https://demo.allennlp.org/next-token-lm?text=AllenNLP%20is

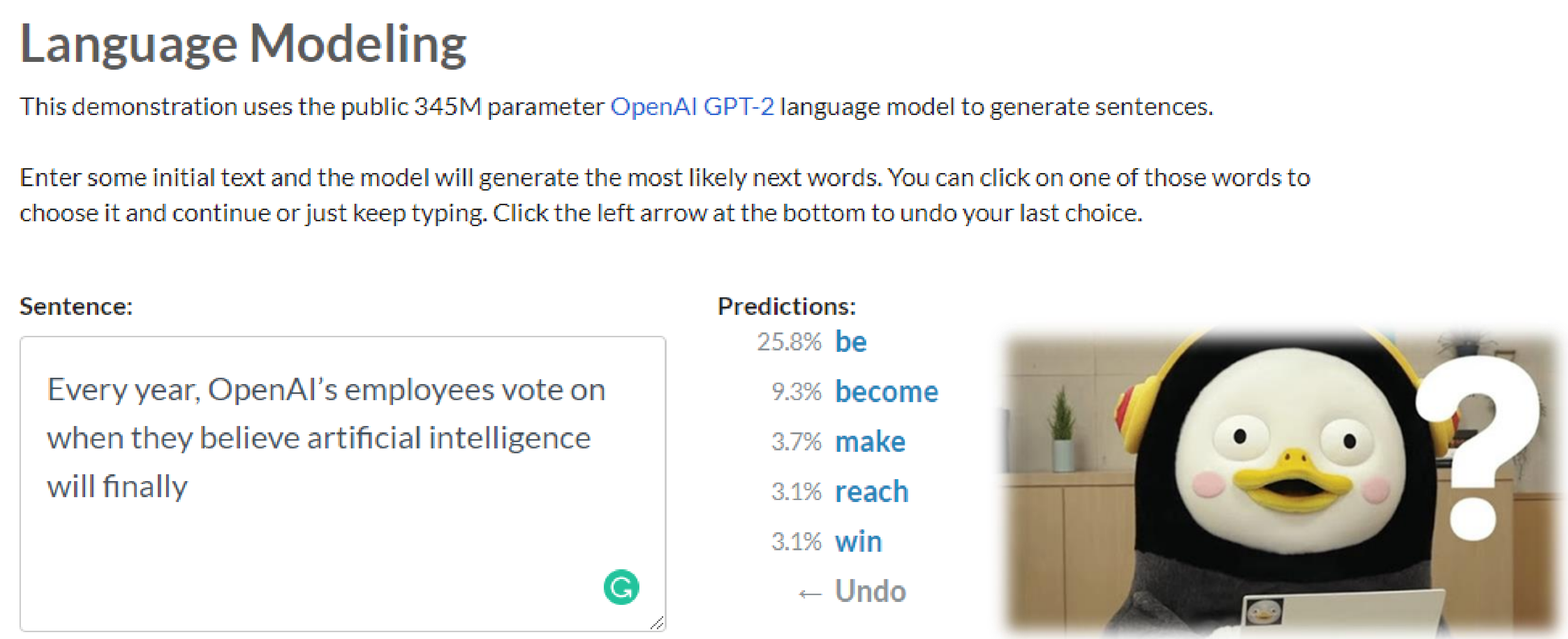
GPT-2: Language Models are Unsupervised Multitask Learners

Radford et. al (2019)

- GPT-2 Example



Every year, OpenAI's employees vote on when they believe artificial general intelligence, or AGI, will finally arrive. It's mostly seen as a fun way to bond, and their estimates differ widely. But in a field that still debates whether human-like autonomous systems are even possible, half the lab bets it is likely to happen within 15 years.



Language Modeling

This demonstration uses the public 345M parameter OpenAI GPT-2 language model to generate sentences.

Enter some initial text and the model will generate the most likely next words. You can click on one of those words to choose it and continue or just keep typing. Click the left arrow at the bottom to undo your last choice.

Sentence:

Every year, OpenAI's employees vote on when they believe artificial intelligence will finally

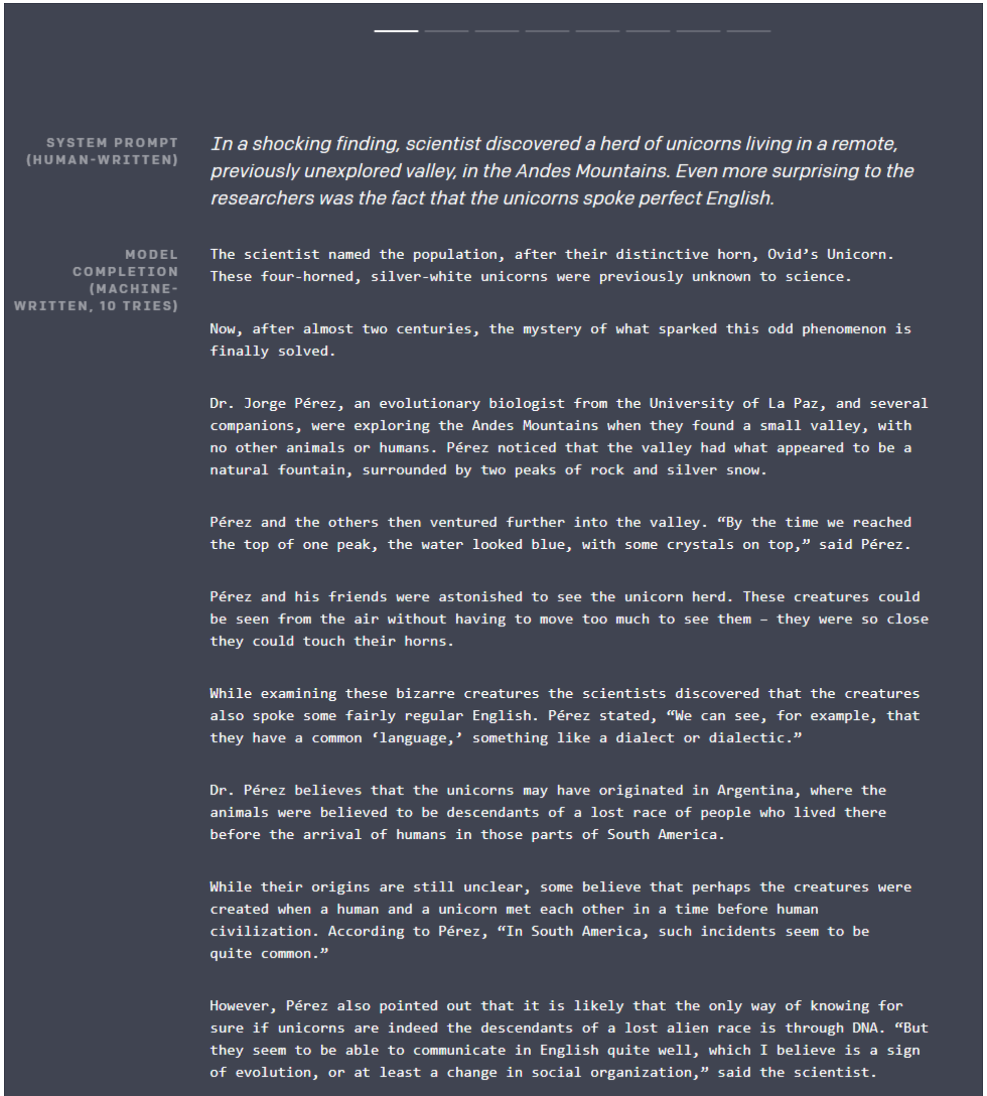
Predictions:

25.8%	be
9.3%	become
3.7%	make
3.1%	reach
3.1%	win
←	Undo

GPT-2: Language Models are Unsupervised Multitask Learners

Radford et. al (2019)

- 오늘의 교훈



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."

However, Pérez also pointed out that it is likely that the only way of knowing for sure if unicorns are indeed the descendants of a lost alien race is through DNA. "But they seem to be able to communicate in English quite well, which I believe is a sign of evolution, or at least a change in social organization," said the scientist.

CEO들에게 이런 거 보여주지 말자

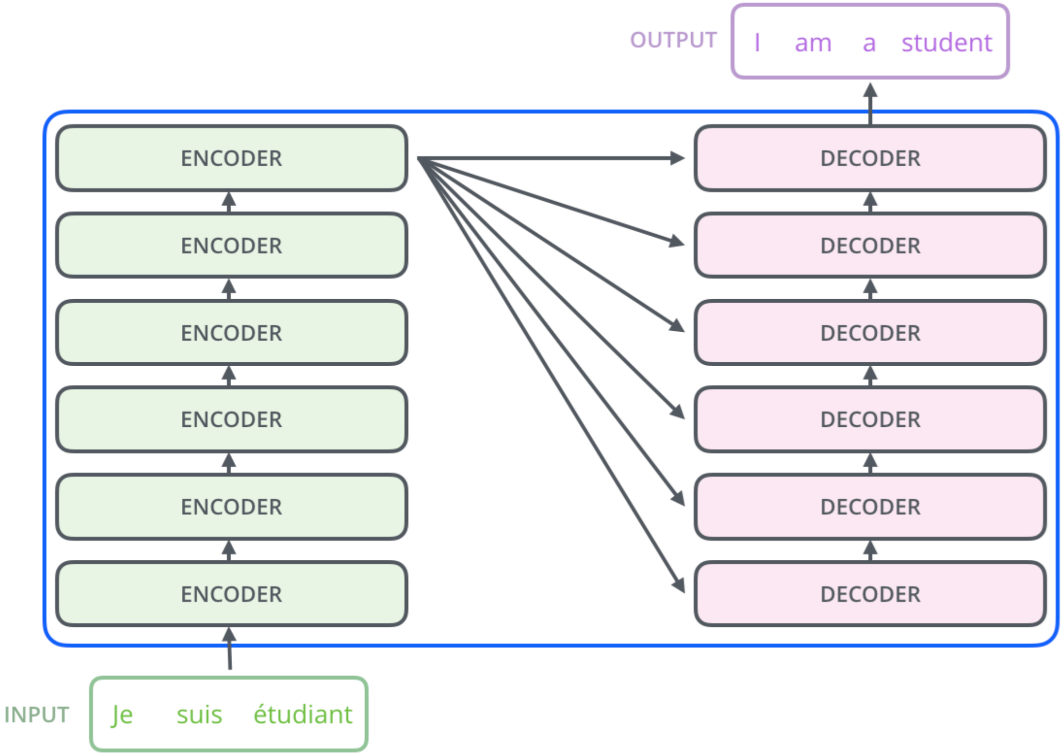
겁나 쉬운 줄 안다

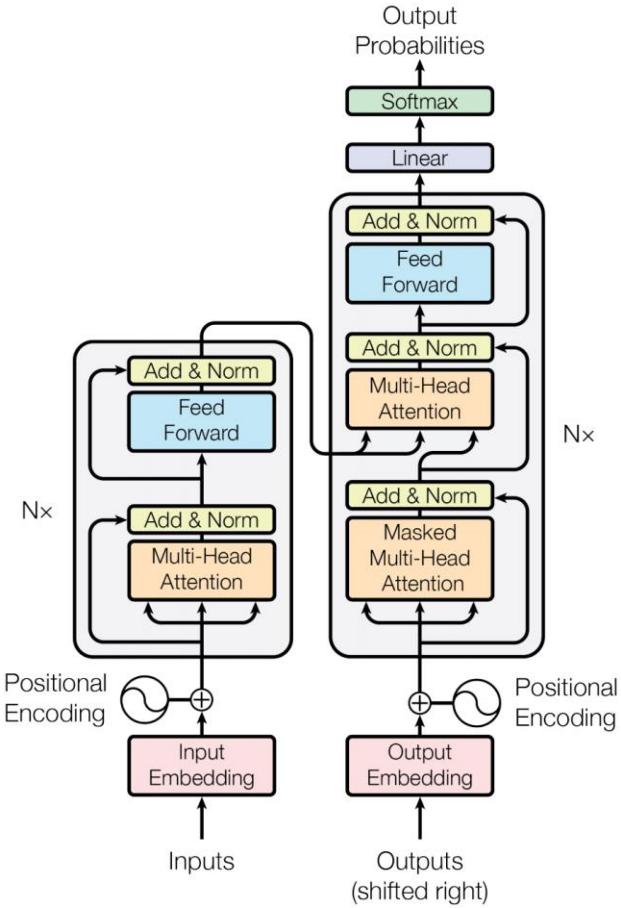
그리고 당신은 곧…

GPT-2: Language Models are Unsupervised Multitask Learners

Alammar (GPT-2)

- Transformer revisited

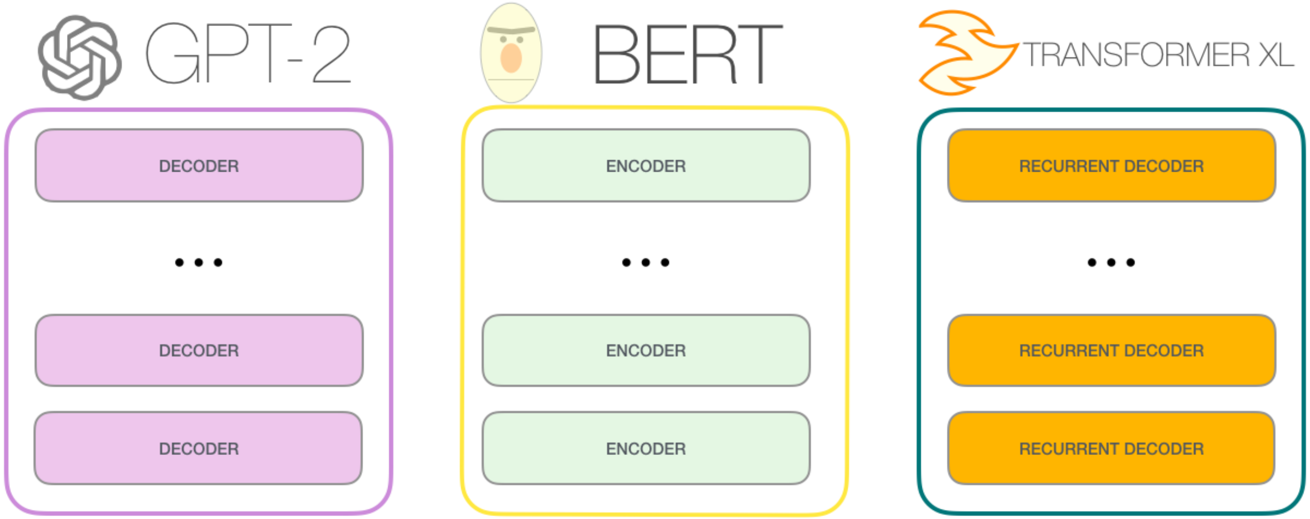




GPT-2: Language Models are Unsupervised Multitask Learners

Alammar (GPT-2)

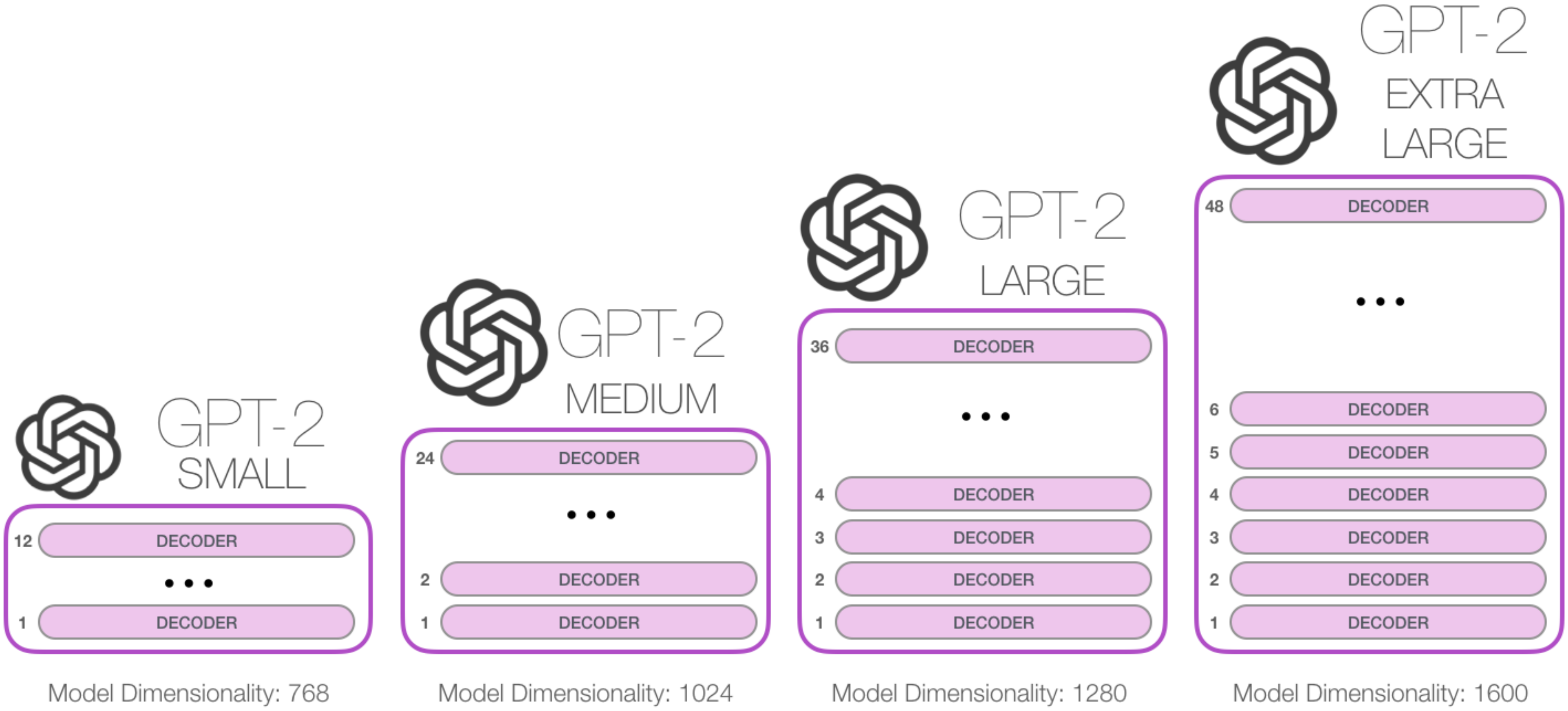
- Transformer revisited
  - ✓ A lot of the subsequent research work saw the architecture shed either the encoder or decoder, and use just one stack of transformer blocks
  - ✓ Stacking them up as high as practically possible, feeding them massive amounts of training text, and throwing vast amounts of compute at them
  - ✓ Hundreds of thousands of dollars to train some of these language models, likely millions in the case of AlphaStar



GPT-2: Language Models are Unsupervised Multitask Learners

Alammar (GPT-2)

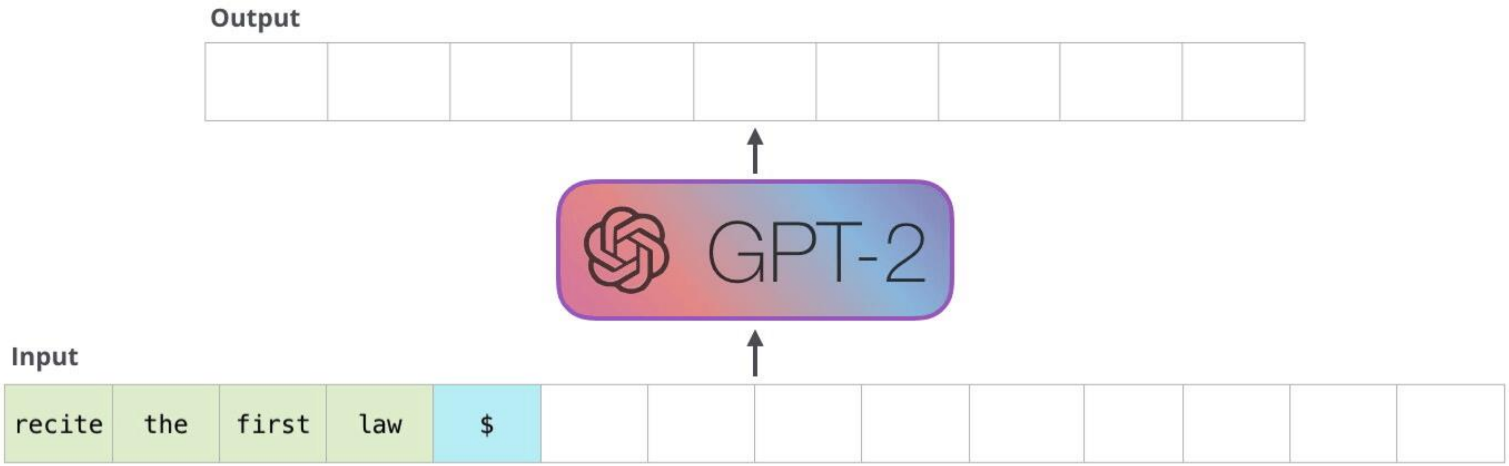
- How high can we stack up the blocks?
  - ✓ Main distinguishing factor of different GPT-2 models



GPT-2: Language Models are Unsupervised Multitask Learners

Alammar (GPT-2)

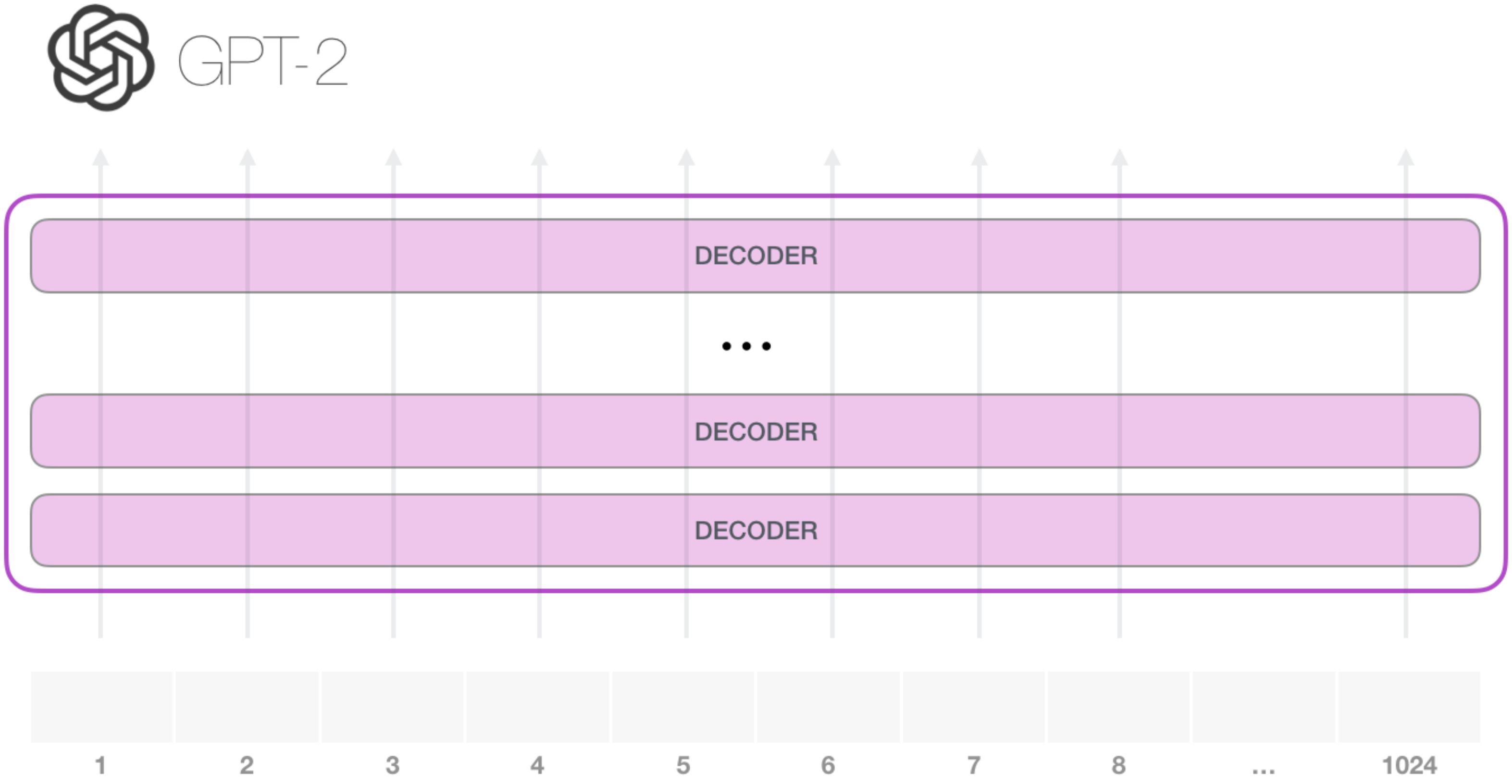
- Main difference between GPT-2 and BERT
  - ✓ GPT-2 is auto-regressive but BERT is not
    - After each token is produced, that token is added to the sequence of inputs



GPT-2: Language Models are Unsupervised Multitask Learners

Alammar (GPT-2)

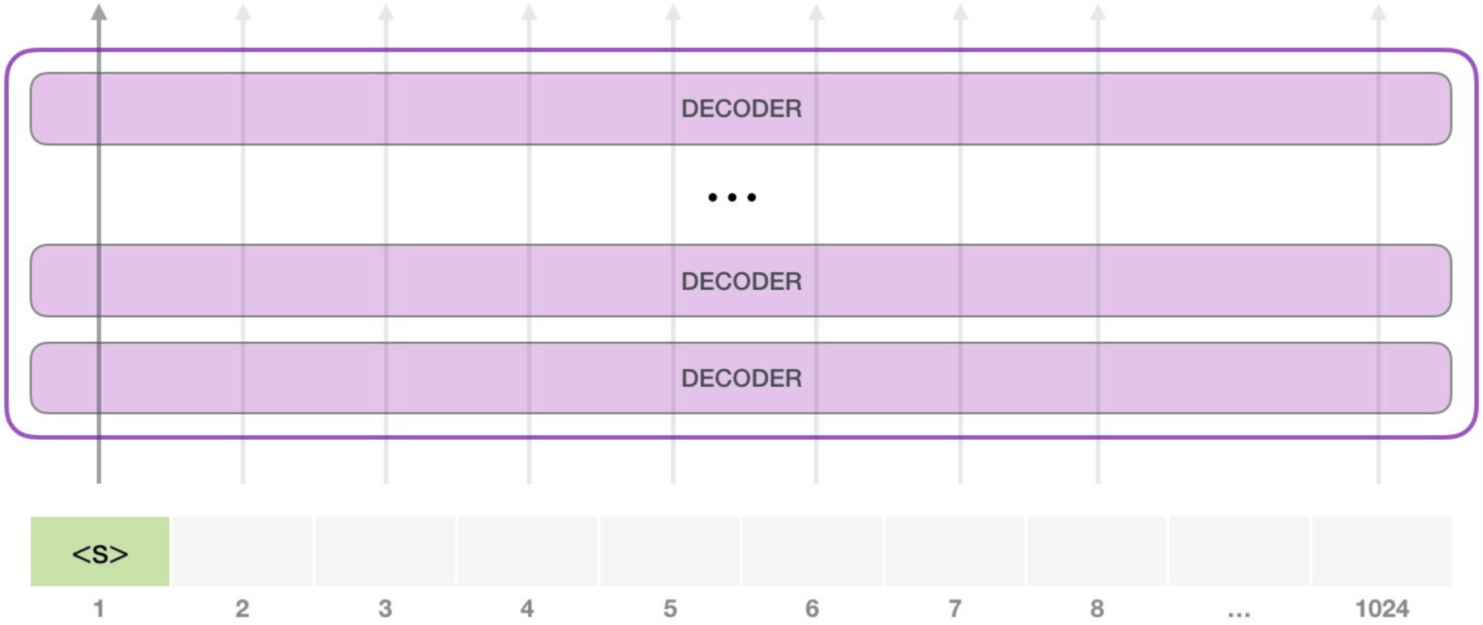
- GPT-2 can process 1024 tokens
  - ✓ Each token flows through all the decoder blocks along its own path



GPT-2: Language Models are Unsupervised Multitask Learners

Alammar (GPT-2)

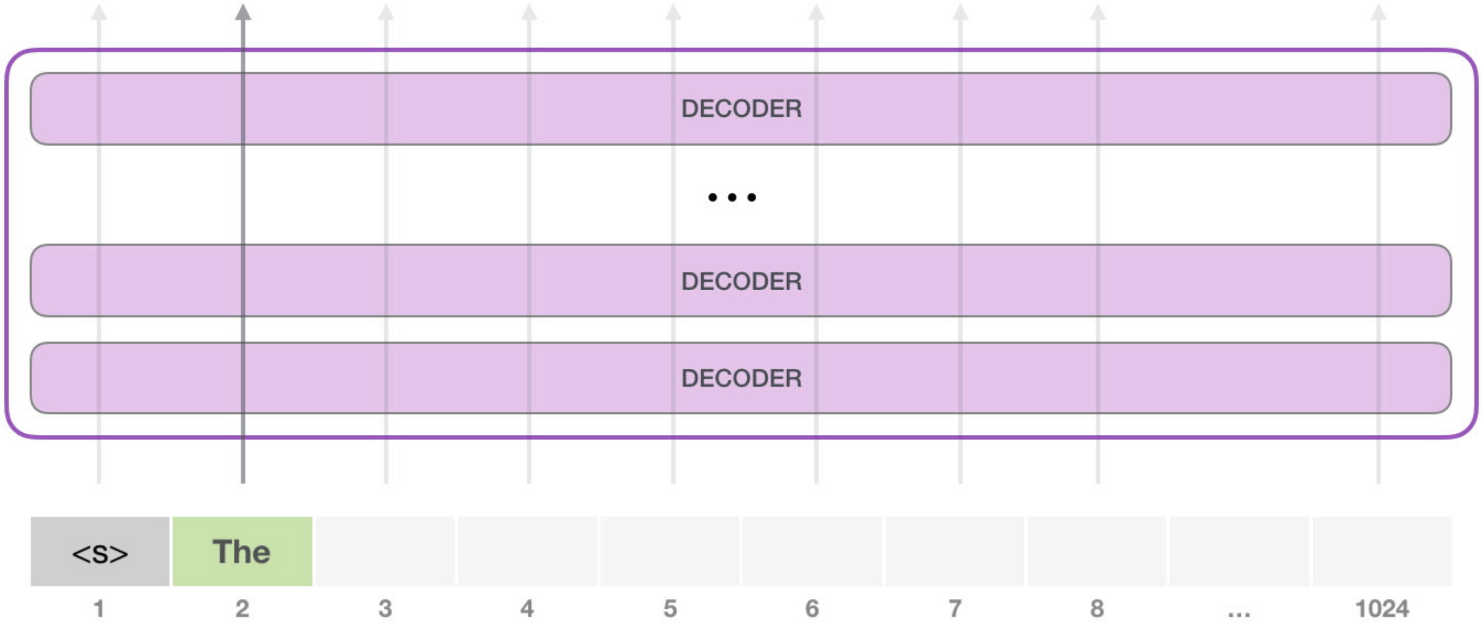
- The simplest way to run a trained GPT-2 is to allow it to ramble on its own
  - ✓ Generating unconditional samples
  - ✓ GPT-2 has a parameter called top-k that we can use to have the model consider sampling words other than the top word



GPT-2: Language Models are Unsupervised Multitask Learners

Alammar (GPT-2)

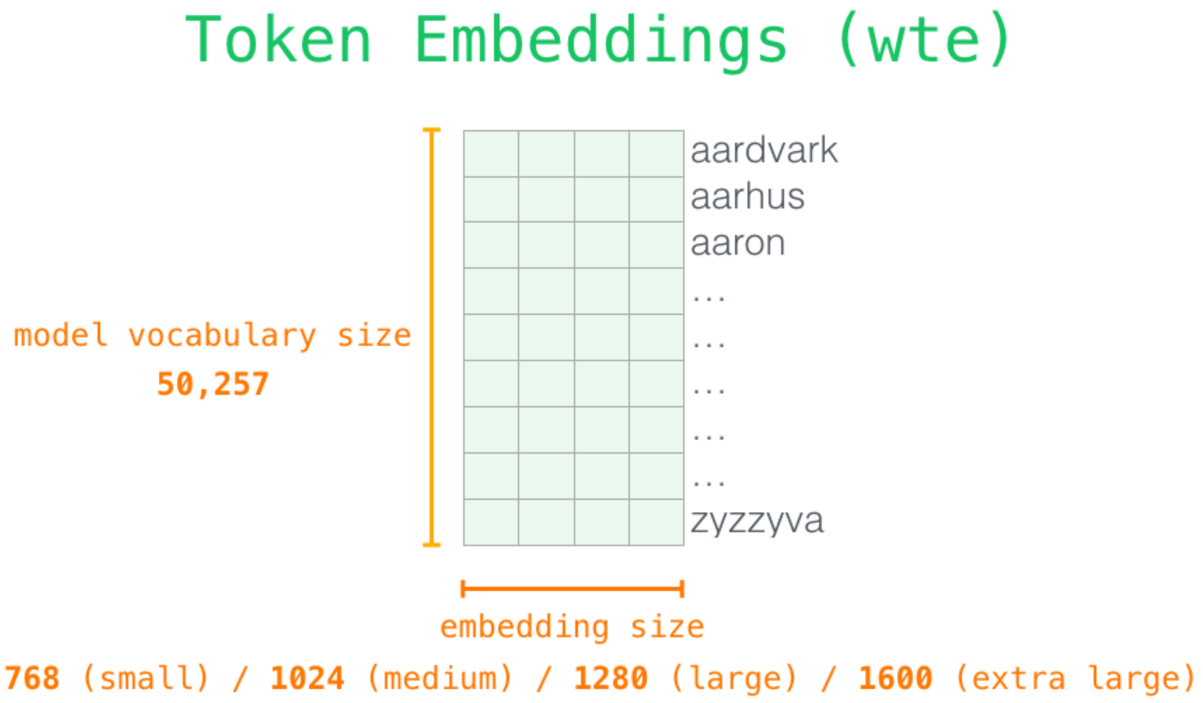
- In the next step, we add the output from the first step to our input sequence, and have the model make its next prediction:
  - ✓ The second path is the only that's active in this calculation
  - ✓ GPT-2 does not re-interpret the first token in light of the second token



GPT-2: Language Models are Unsupervised Multitask Learners

Alammar (GPT-2)

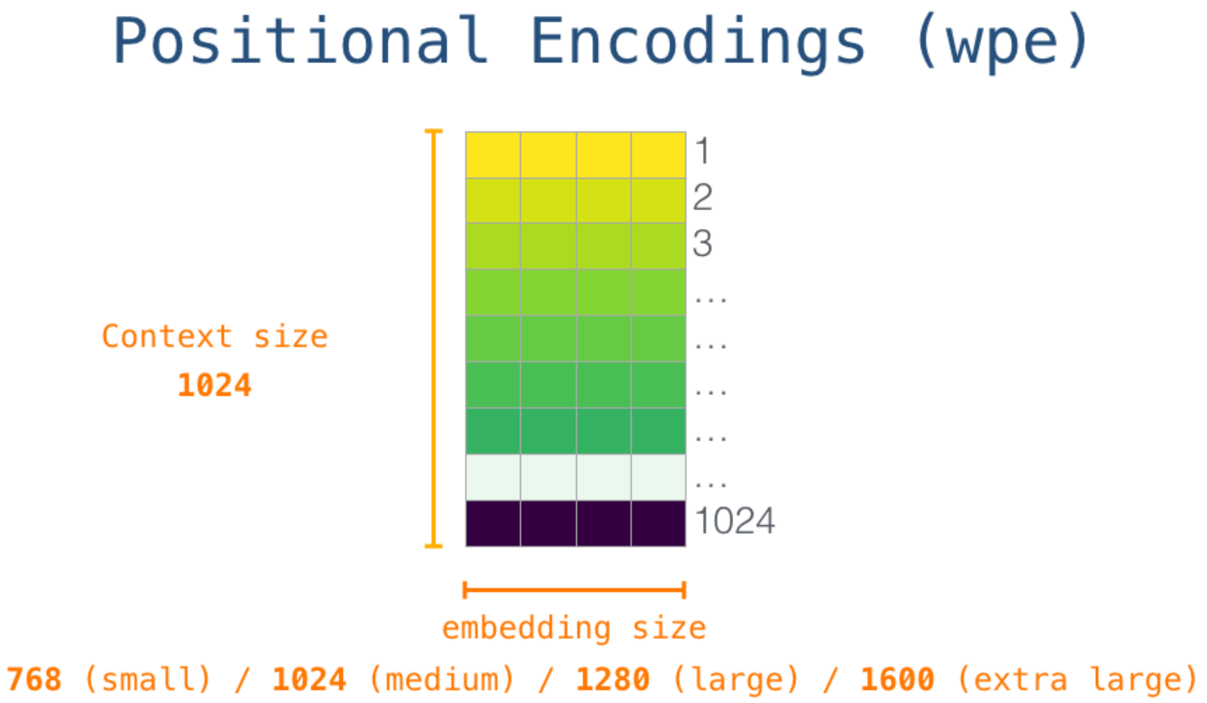
- GPT2: A deeper look inside
  - ✓ Input encoding



GPT-2: Language Models are Unsupervised Multitask Learners

Alammar (GPT-2)

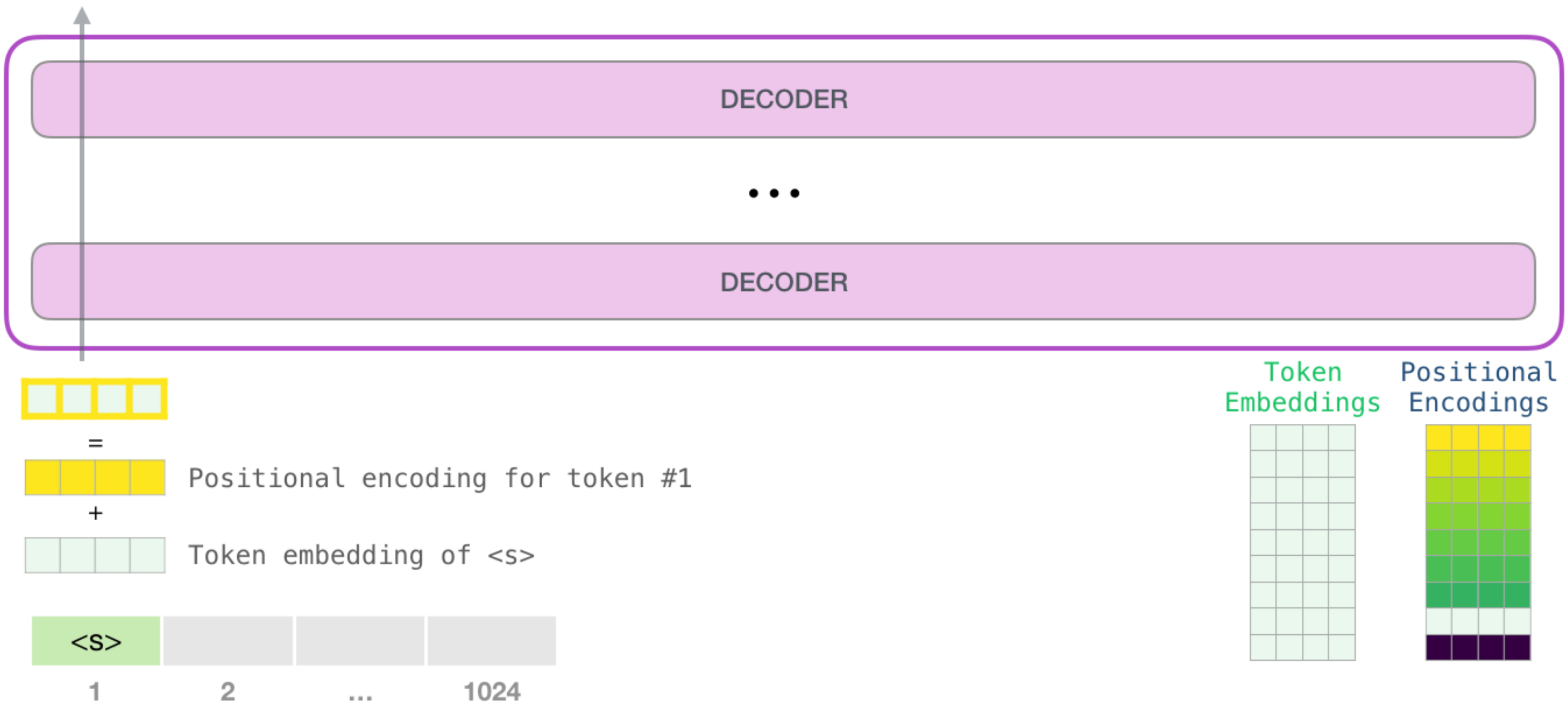
- GPT2: A deeper look inside
  - ✓ Positional encoding


Positional Encodings (wpe)
Context size
1024
1
2
3
...
...
...
...
...
1024
embedding size
768 (small) / 1024 (medium) / 1280 (large) / 1600 (extra large)

GPT-2: Language Models are Unsupervised Multitask Learners

Alammar (GPT-2)

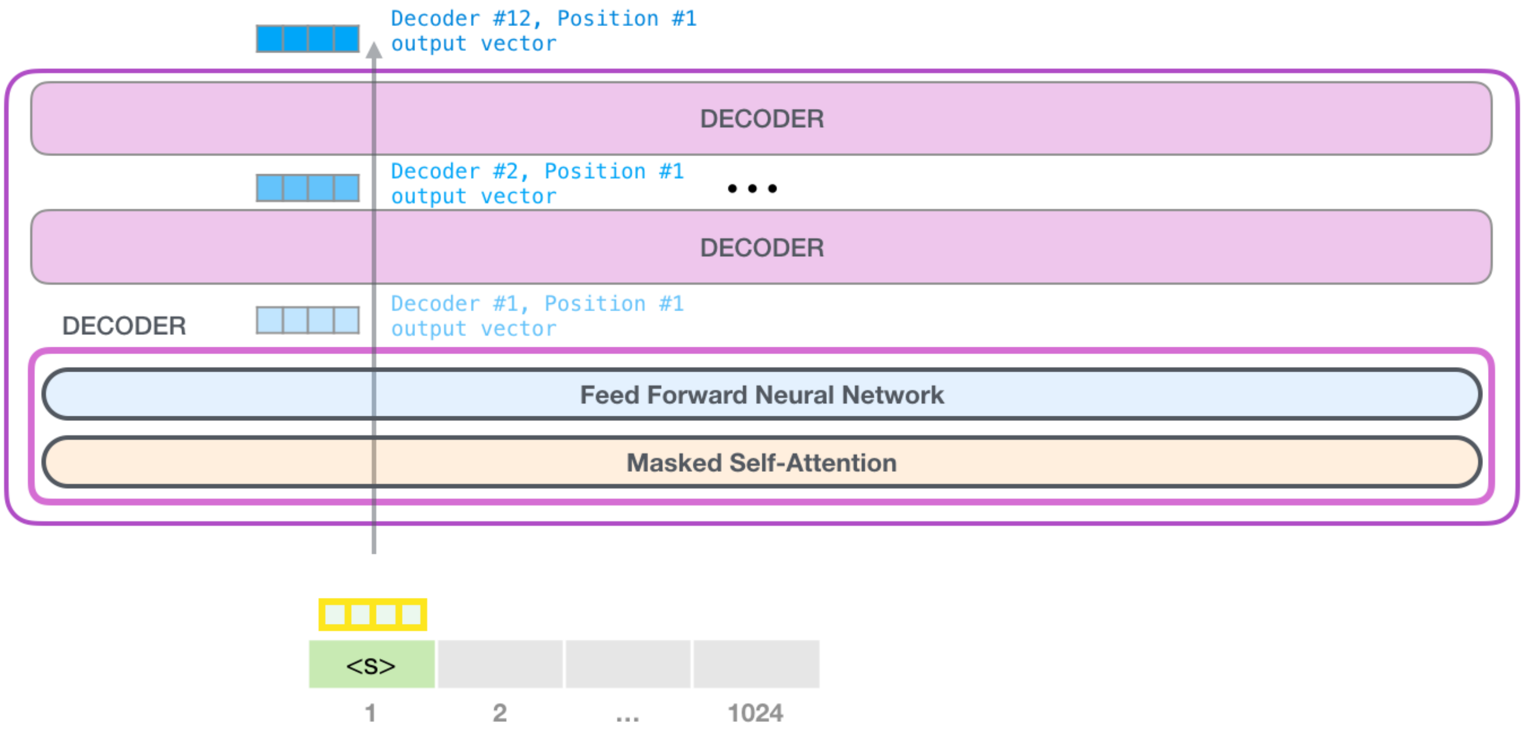
- GPT2: A deeper look inside
  - ✓ Sending a word to the first transformer block



GPT-2: Language Models are Unsupervised Multitask Learners

Alammar (GPT-2)

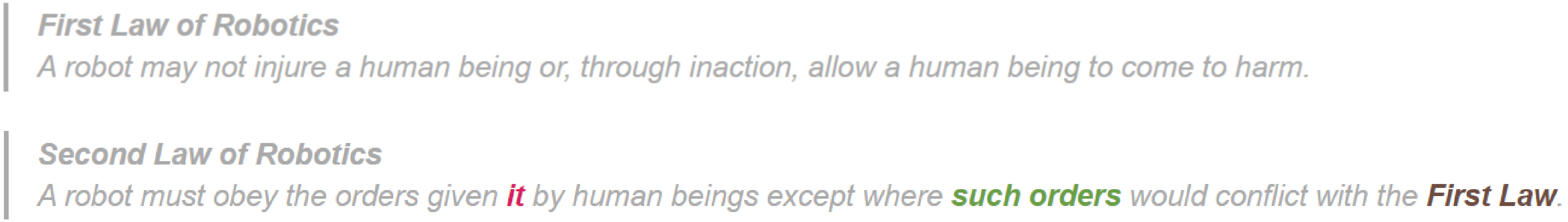
- A journey up the stack
  - ✓ Once a lower-level transformer block processes the token, it sends its resulting vector up the stack to be processed by the next block
    - The process is identical in each block, but each block has its own weights in both self-attention and the neural network sublayers



GPT-2: Language Models are Unsupervised Multitask Learners

Alammar (GPT-2)

- Self-Attention Recap
  - ✓ Language heavily relies on context
    - ▪ Look at the second law



> ***First Law of Robotics***
> *A robot may not injure a human being or, through inaction, allow a human being to come to harm.*

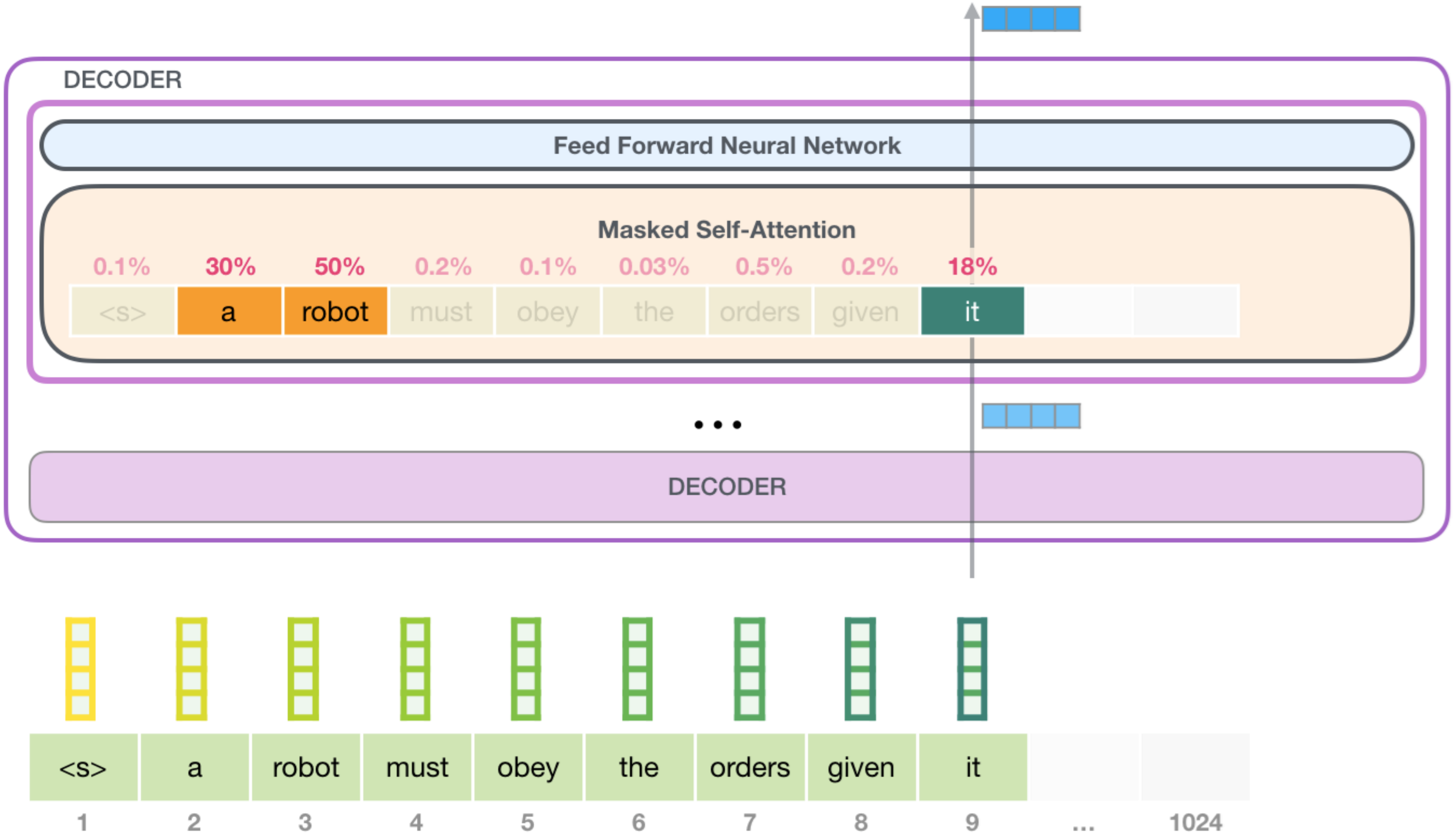
> ***Second Law of Robotics***
> *A robot must obey the orders given **it** by human beings except where **such orders** would conflict with the **First Law**.*

  - ✓ When a model processes this sentence, it has to be able to know that
    - ▪ *it* refers to the robot
    - ▪ *such orders* refers to the earlier part of the law, namely "the orders given it by human beings"
    - ▪ *The First Law* refers to the entire First Law
  - ✓ This is what self-attention does

GPT-2: Language Models are Unsupervised Multitask Learners

Alammar (GPT-2)

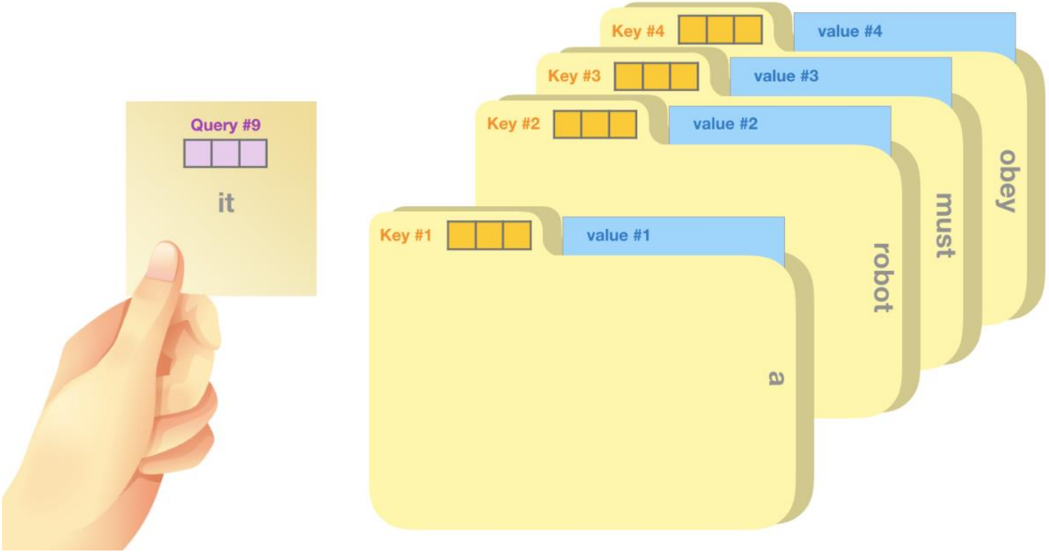
- Self-Attention Recap
  - ✓ This self-attention layer in the top block is paying attention to "a robot" when it processes the word "it"



GPT-2: Language Models are Unsupervised Multitask Learners

Alammar (GPT-2)

- Self-Attention Recap
  - ✓ Think of it like searching through a filing cabinet

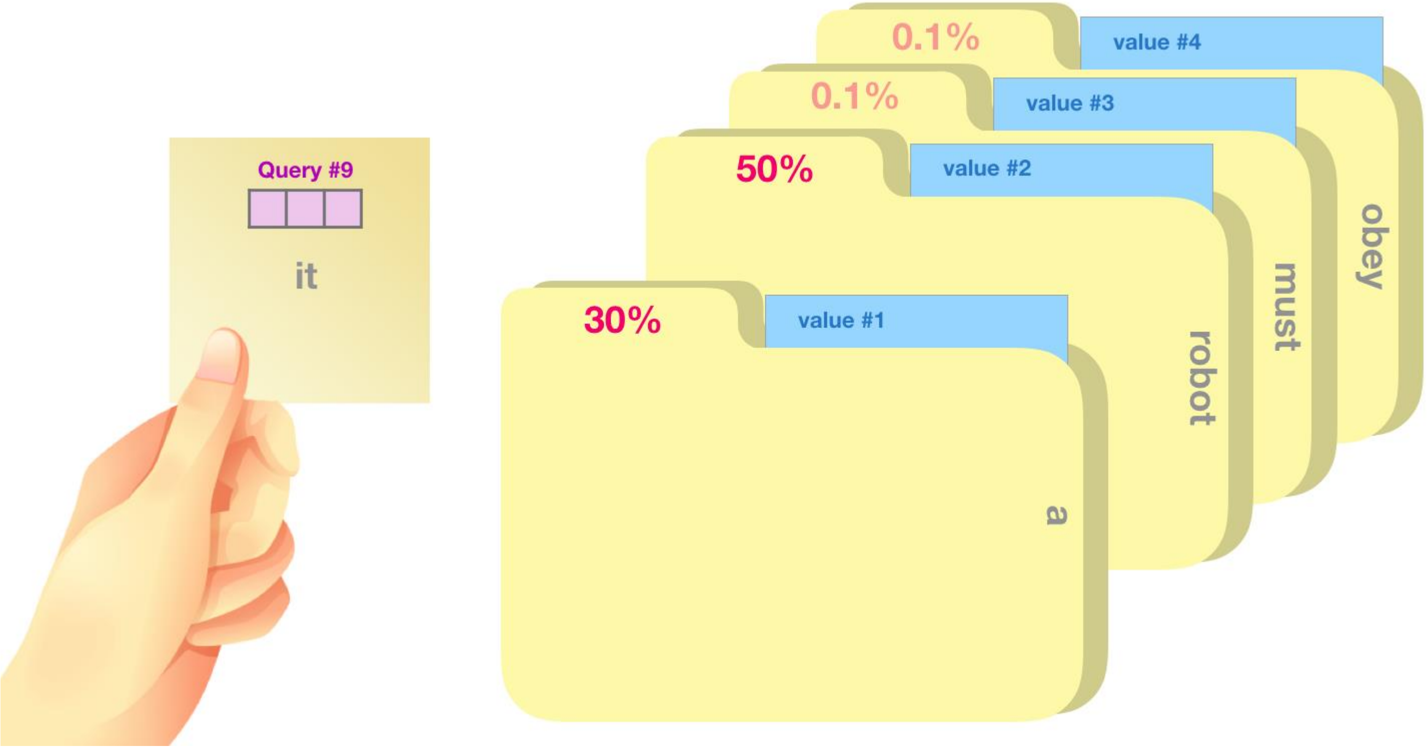


    - The query is like a sticky note with the topic you're researching
    - The keys are like the labels of the folders inside the cabinet
    - When you match the tag with a sticky note, we take out the contents of that folder, these contents are the value vector
    - Except you're not only looking for one value, but a blend of values from a blend of folders.

GPT-2: Language Models are Unsupervised Multitask Learners

Alammar (GPT-2)

- Self-Attention Recap
  - ✓ Multiplying the query vector by each key vector produces a score for each folder (technically: dot product followed by softmax)

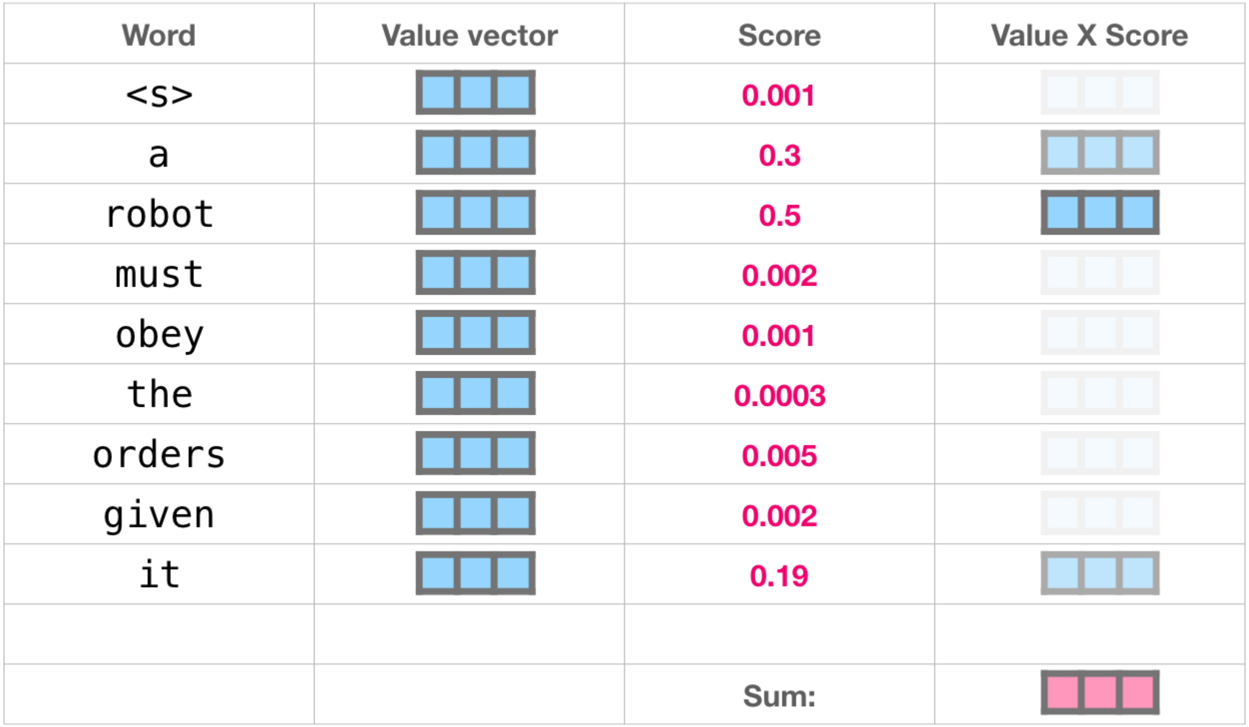


GPT-2: Language Models are Unsupervised Multitask Learners

Alammar (GPT-2)

- Self-Attention Recap

  ✓ Multiply each value by its score and sum up – resulting in our self-attention outcome



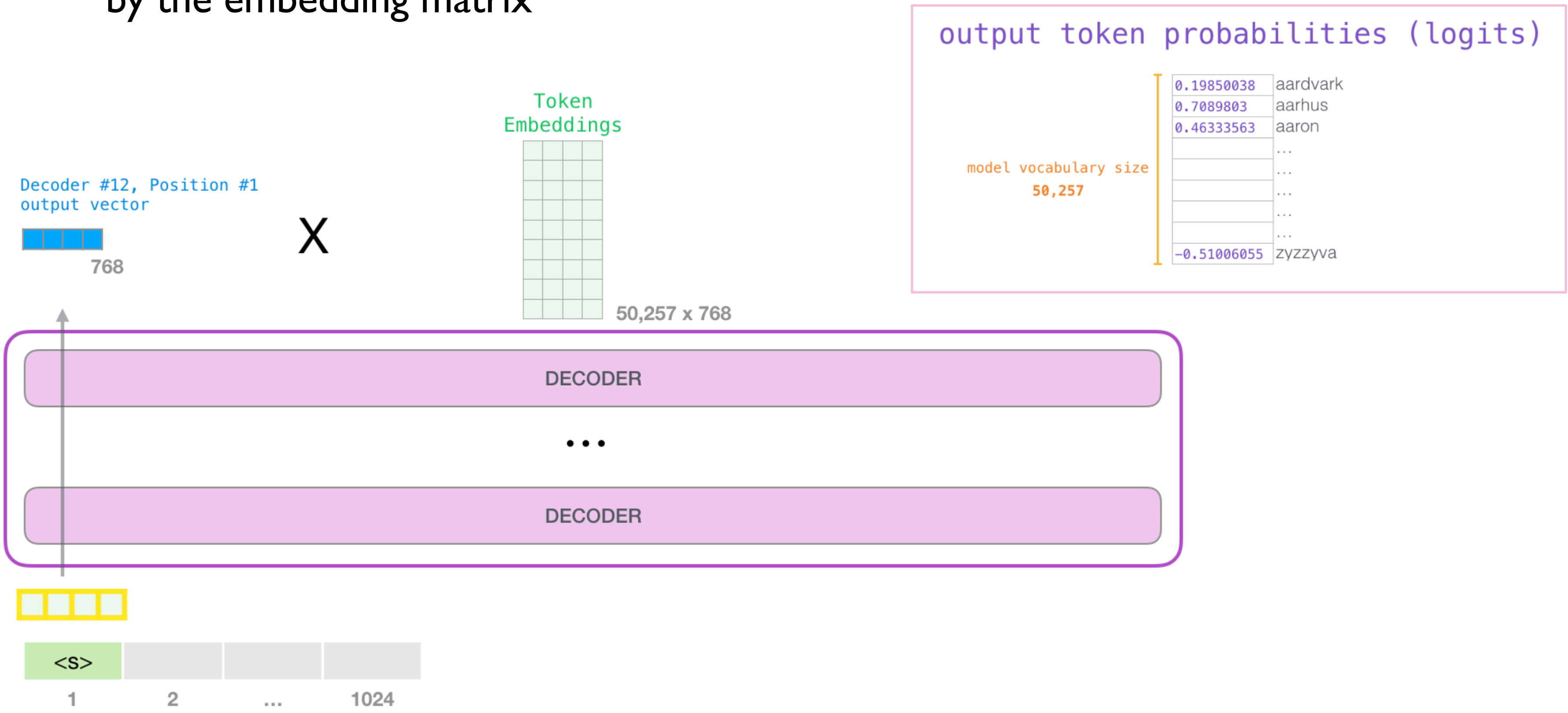
Word	Value vector	Score	Value X Score
<s>	🟦🟦🟦	0.001	□□□
a	🟦🟦🟦	0.3	🟦🟦🟦
robot	🟦🟦🟦	0.5	🟦🟦🟦
must	🟦🟦🟦	0.002	□□□
obey	🟦🟦🟦	0.001	□□□
the	🟦🟦🟦	0.0003	□□□
orders	🟦🟦🟦	0.005	□□□
given	🟦🟦🟦	0.002	□□□
it	🟦🟦🟦	0.19	🟦🟦🟦
		Sum:	🟥🟥🟥

  - This weighted blend of value vectors results in a vector that paid 50% of its attention to the word `robot`, 30% to the word `a`, and 19% to the word `it`

GPT-2: Language Models are Unsupervised Multitask Learners

Alammar (GPT-2)

- Model Output
  - ✓ When the top block in the model produces its output vector (the result of its own self-attention followed by its own neural network), the model multiplies that vector by the embedding matrix

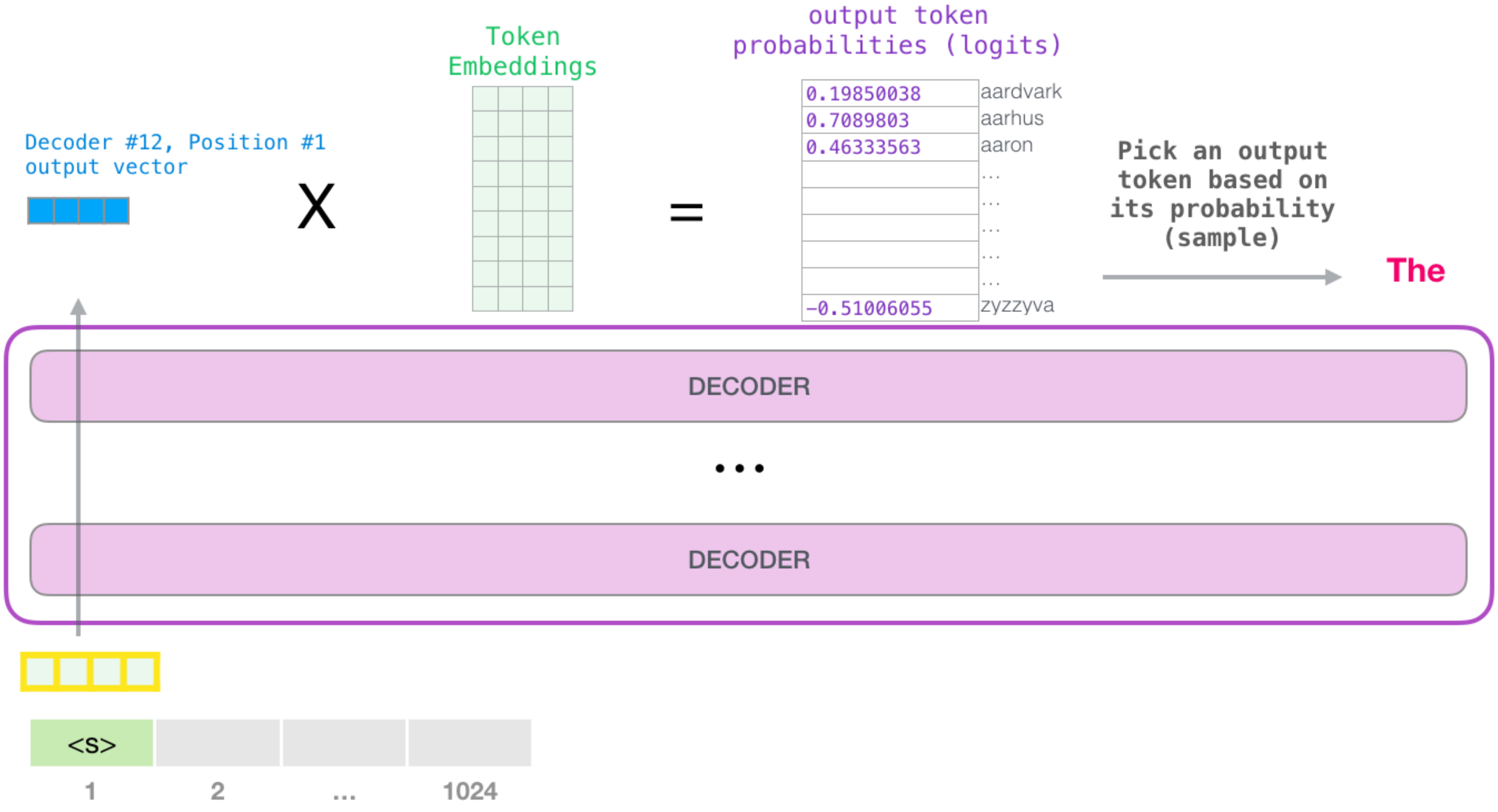


0.19850038	aardvark
0.7089803	aarhus
0.46333563	aaron
	...
	...
	...
	...
	...
-0.51006055	zyzzyva

GPT-2: Language Models are Unsupervised Multitask Learners

Alammar (GPT-2)

- Model Output



GPT-2: Language Models are Unsupervised Multitask Learners

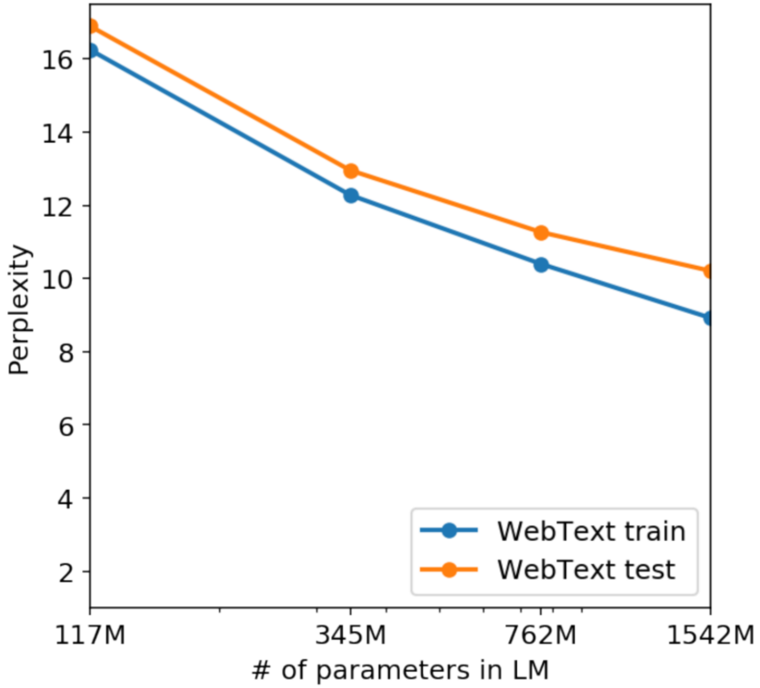
Alammar (GPT-2)

- Notes in GPT-2
  - ✓ GPT-2 uses Byte Pair Encoding to create the tokens in its vocabulary; tokens are usually parts of words
  - ✓ When training, a maximum of 512 tokens are processes at the same time
  - ✓ Layer normalization is important in Transformer structure

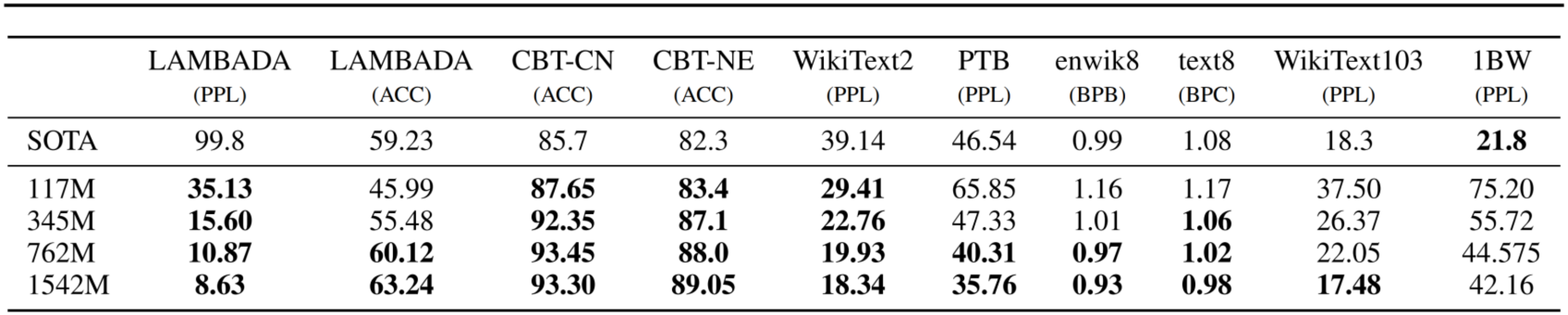
GPT-2: Language Models are Unsupervised Multitask Learners

Alammar (GPT-2)

- Experiments
  - ✓ Performance w.r.t. model size



  - ✓ Zero-shot results on Language Modeling datasets

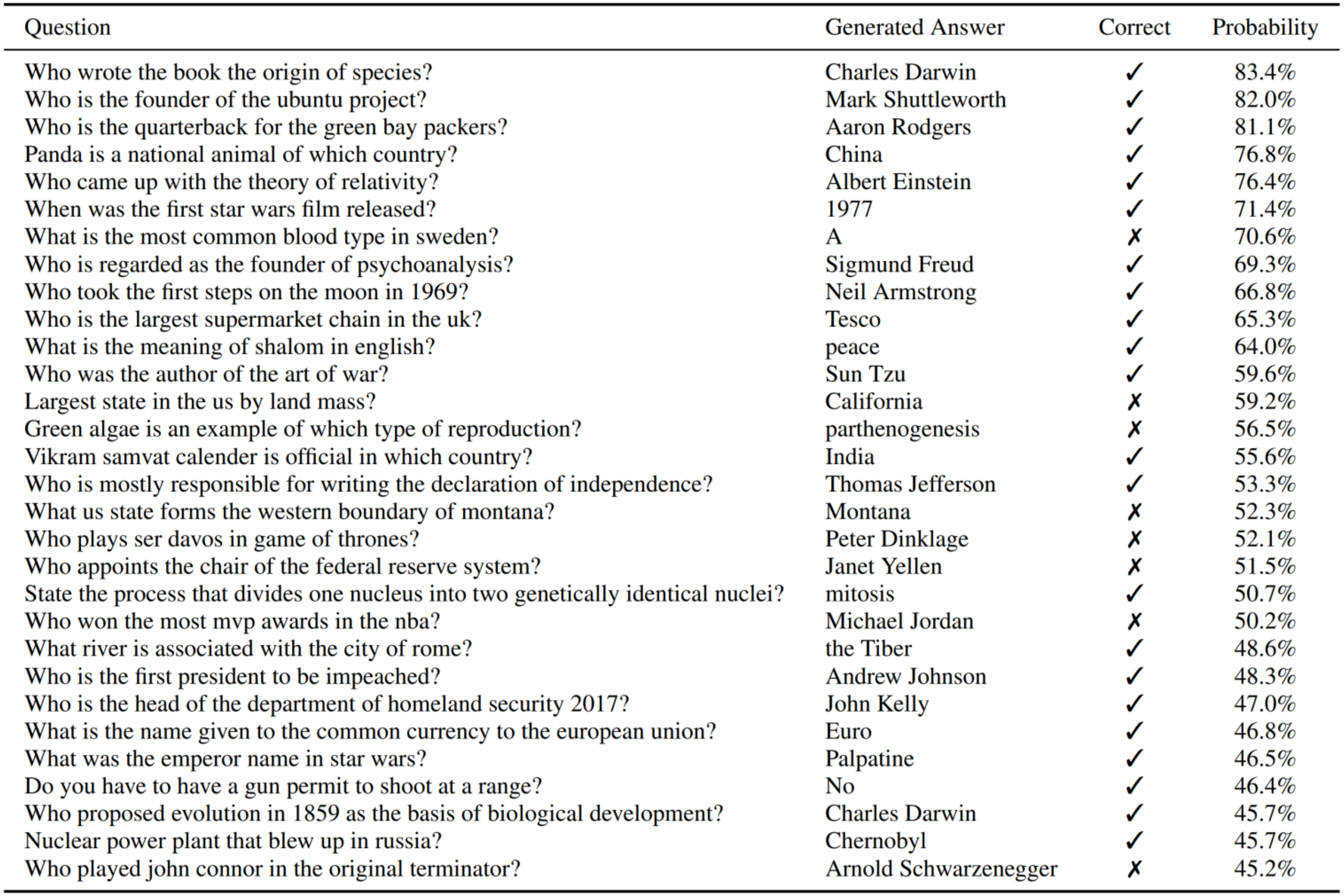


	LAMBADA (PPL)	LAMBADA (ACC)	CBT-CN (ACC)	CBT-NE (ACC)	WikiText2 (PPL)	PTB (PPL)	enwik8 (BPB)	text8 (BPC)	WikiText103 (PPL)	1BW (PPL)
SOTA	99.8	59.23	85.7	82.3	39.14	46.54	0.99	1.08	18.3	**21.8**
117M	**35.13**	45.99	**87.65**	**83.4**	**29.41**	65.85	1.16	1.17	37.50	75.20
345M	**15.60**	55.48	**92.35**	**87.1**	**22.76**	47.33	1.01	**1.06**	26.37	55.72
762M	**10.87**	**60.12**	**93.45**	**88.0**	**19.93**	**40.31**	**0.97**	**1.02**	22.05	44.575
1542M	**8.63**	**63.24**	**93.30**	**89.05**	**18.34**	**35.76**	**0.93**	**0.98**	**17.48**	42.16

GPT-2: Language Models are Unsupervised Multitask Learners

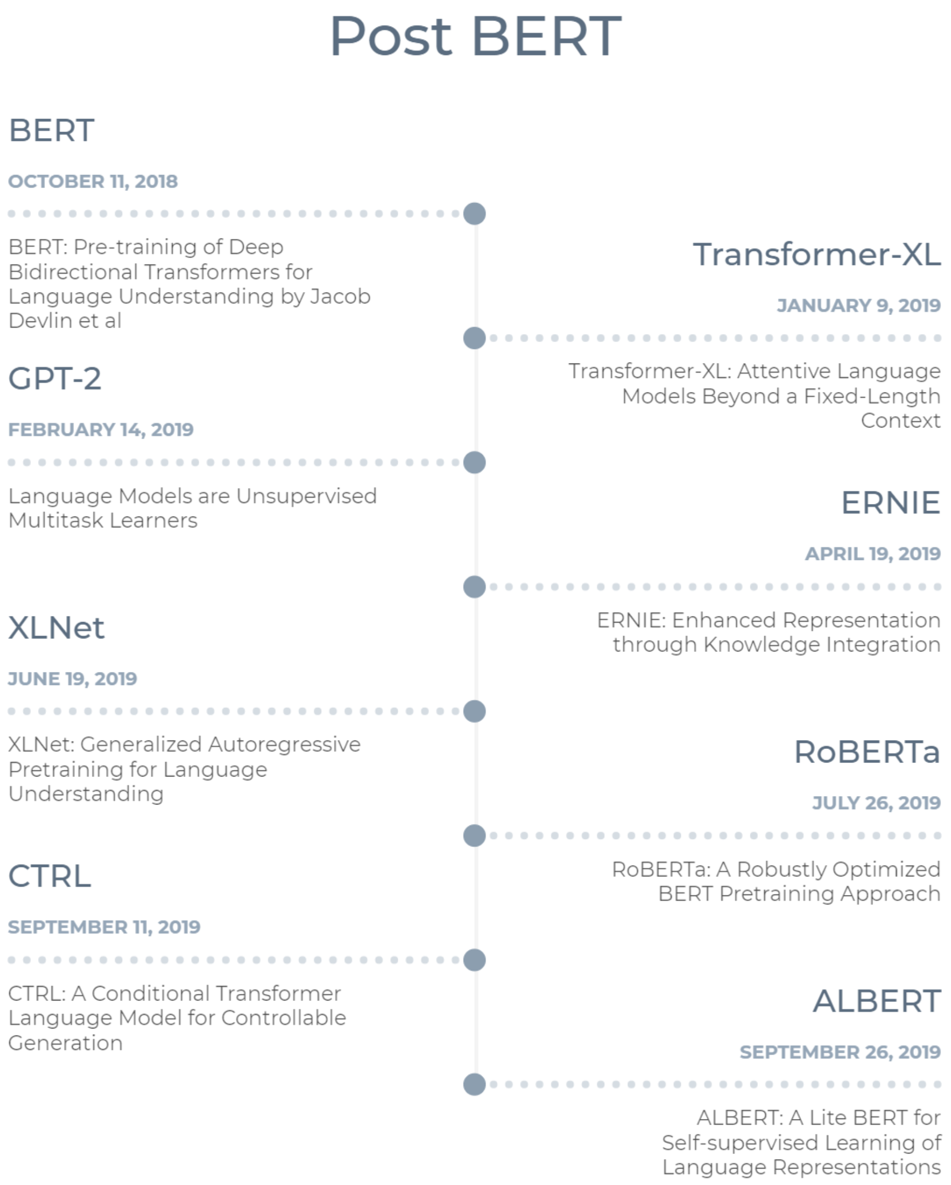
Alammar (GPT-2)

- Experiments
  - ✓ Answers for the questions not in the training dataset



Question	Generated Answer	Correct	Probability
Who wrote the book the origin of species?	Charles Darwin	✓	83.4%
Who is the founder of the ubuntu project?	Mark Shuttleworth	✓	82.0%
Who is the quarterback for the green bay packers?	Aaron Rodgers	✓	81.1%
Panda is a national animal of which country?	China	✓	76.8%
Who came up with the theory of relativity?	Albert Einstein	✓	76.4%
When was the first star wars film released?	1977	✓	71.4%
What is the most common blood type in sweden?	A	✗	70.6%
Who is regarded as the founder of psychoanalysis?	Sigmund Freud	✓	69.3%
Who took the first steps on the moon in 1969?	Neil Armstrong	✓	66.8%
Who is the largest supermarket chain in the uk?	Tesco	✓	65.3%
What is the meaning of shalom in english?	peace	✓	64.0%
Who was the author of the art of war?	Sun Tzu	✓	59.6%
Largest state in the us by land mass?	California	✗	59.2%
Green algae is an example of which type of reproduction?	parthenogenesis	✗	56.5%
Vikram samvat calender is official in which country?	India	✓	55.6%
Who is mostly responsible for writing the declaration of independence?	Thomas Jefferson	✓	53.3%
What us state forms the western boundary of montana?	Montana	✗	52.3%
Who plays ser davos in game of thrones?	Peter Dinklage	✗	52.1%
Who appoints the chair of the federal reserve system?	Janet Yellen	✗	51.5%
State the process that divides one nucleus into two genetically identical nuclei?	mitosis	✓	50.7%
Who won the most mvp awards in the nba?	Michael Jordan	✗	50.2%
What river is associated with the city of rome?	the Tiber	✓	48.6%
Who is the first president to be impeached?	Andrew Johnson	✓	48.3%
Who is the head of the department of homeland security 2017?	John Kelly	✓	47.0%
What is the name given to the common currency to the european union?	Euro	✓	46.8%
What was the emperor name in star wars?	Palpatine	✓	46.5%
Do you have to have a gun permit to shoot at a range?	No	✓	46.4%
Who proposed evolution in 1859 as the basis of biological development?	Charles Darwin	✓	45.7%
Nuclear power plant that blew up in russia?	Chernobyl	✓	45.7%
Who played john connor in the original terminator?	Arnold Schwarzenegger	✗	45.2%

Beyond ELMo, GPT, and BERT



Post BERT

BERT

OCTOBER 11, 2018

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding by Jacob Devlin et al

Transformer-XL

JANUARY 9, 2019

Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context

GPT-2

FEBRUARY 14, 2019

Language Models are Unsupervised Multitask Learners

ERNIE

APRIL 19, 2019

ERNIE: Enhanced Representation through Knowledge Integration

XLNet

JUNE 19, 2019

XLNet: Generalized Autoregressive Pretraining for Language Understanding

RoBERTa

JULY 26, 2019

RoBERTa: A Robustly Optimized BERT Pretraining Approach

CTRL

SEPTEMBER 11, 2019

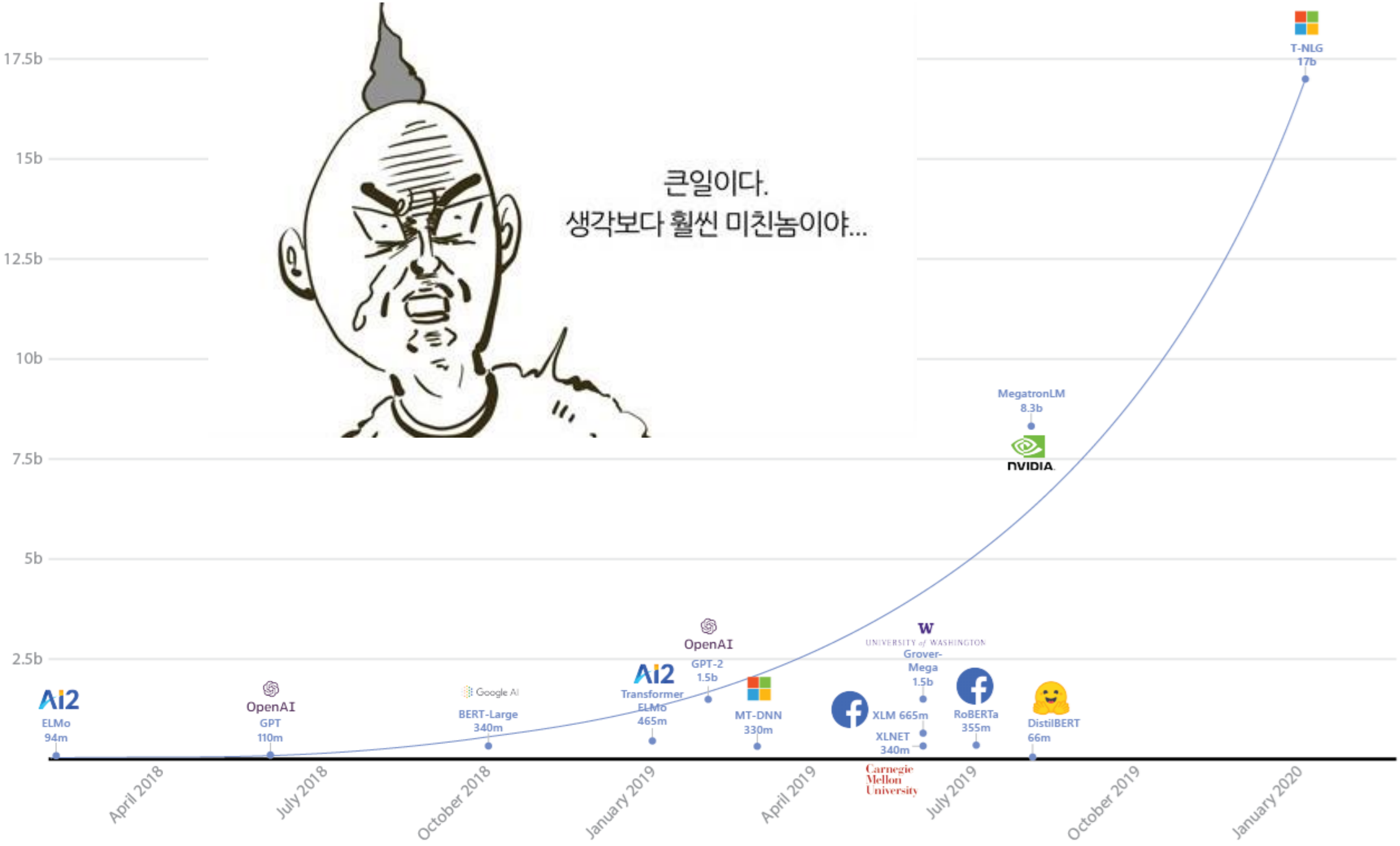
CTRL: A Conditional Transformer Language Model for Controllable Generation

ALBERT

SEPTEMBER 26, 2019

ALBERT: A Lite BERT for Self-supervised Learning of Language Representations

Bigger, BIGger, BIGGER!!



https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/

ANY
questions?