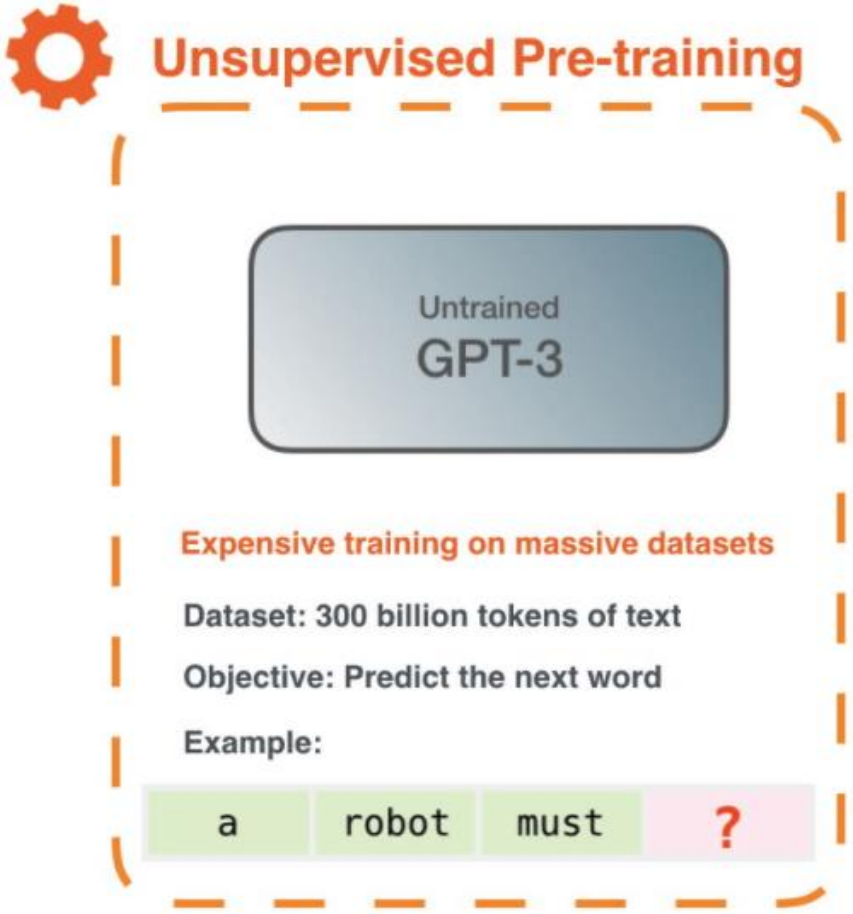
GPT-3

전은주

REFERENCES

- How GPT3 Works - Visualizations and Animations - https://jalammar.github.io/how-gpt3-works-visualizations-animations/
- The Illustrated GPT-2 (Visualizing Transformer Language Models) https://jalammar.github.io/illustrated-gpt2/
- NLP for Developers: GPT-3 | Rasa https://www.youtube.com/watch?v=ZNeNMTSMA5Y
- GPT-3: Language Models are Few-Shot Learners (Paper Explained) https://www.youtube.com/watch?v=SY5PvZrJhLE
- Language Models are Few-Shot Learners https://arxiv.org/pdf/2005.14165.pdf
- [논문리뷰] GPT3 - Language Models are Few-Shot Learners https://littlefoxdiary.tistory.com/44

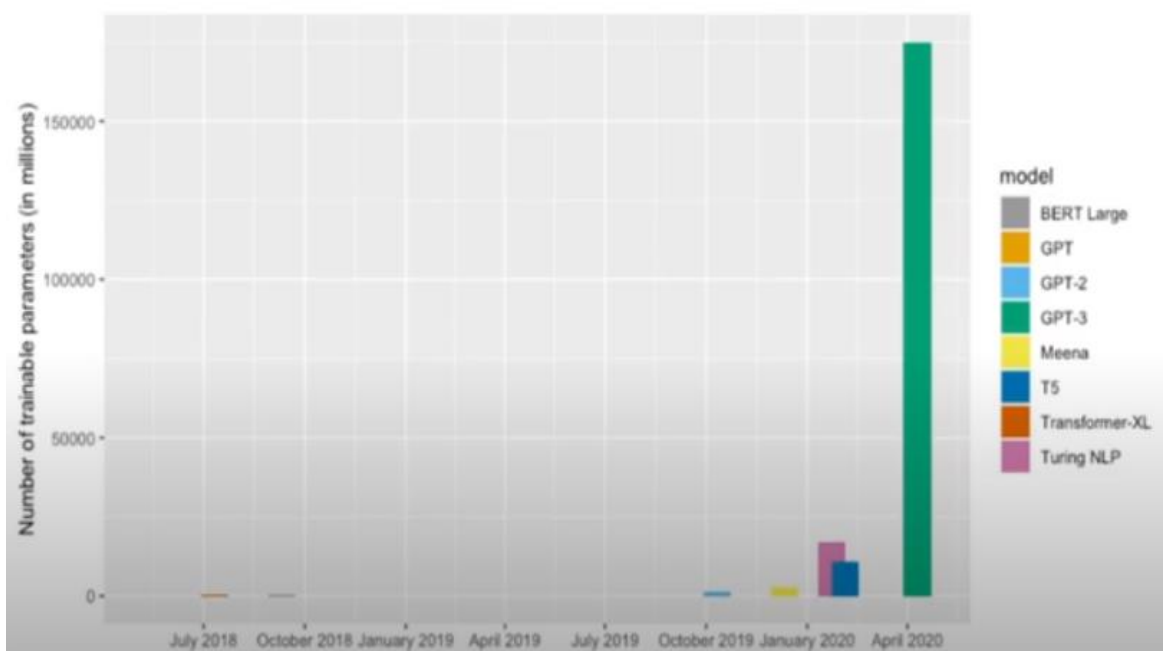
How GPT3 Works



- Language Model: Given input text, it probabilistically **predicts** what tokens from a now vocabulary **will come next**
- The output is generated from what the model “learned” during its training period where it **scanned** vast amounts of text.
- It was estimated to cost 355 GPU years and cost $4.6m($ 4,600,000, 54억 7,170만 원)

Parameters

- GPT3 encodes what it learns from training in 175 billion numbers (1750억, called parameters).
- GPT3 occur inside its stack of 96 transformer decoder layers. Each of these layers has its own 1.8B parameter (18억)



https://www.youtube.com/watch?v=ZNeNMTSMA5Y

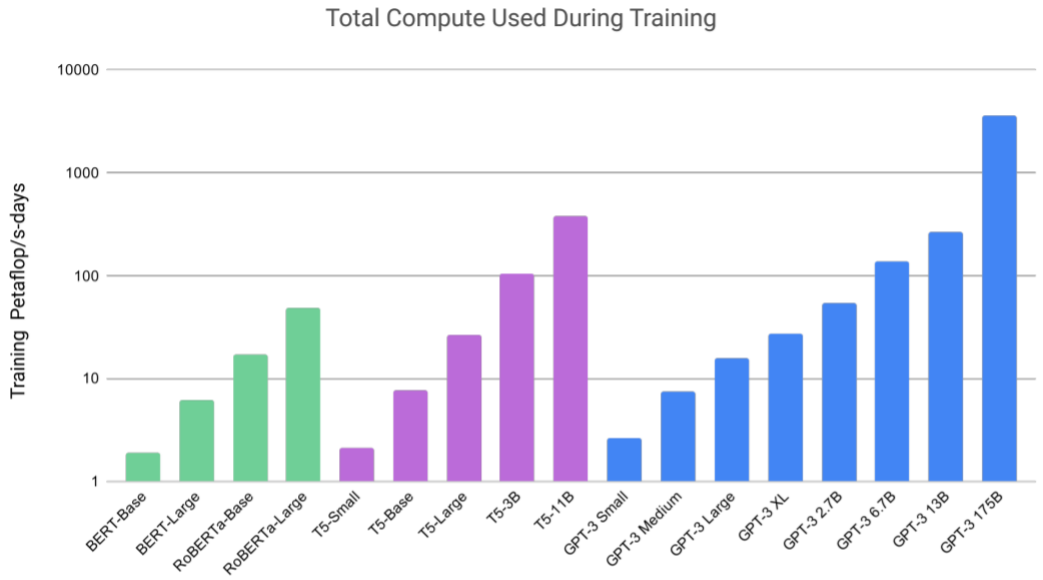


Figure 2.2: Total compute used during training. Based on the analysis in Scaling Laws For Neural Language Models

https://arxiv.org/pdf/2005.14165.pdf

Datasets

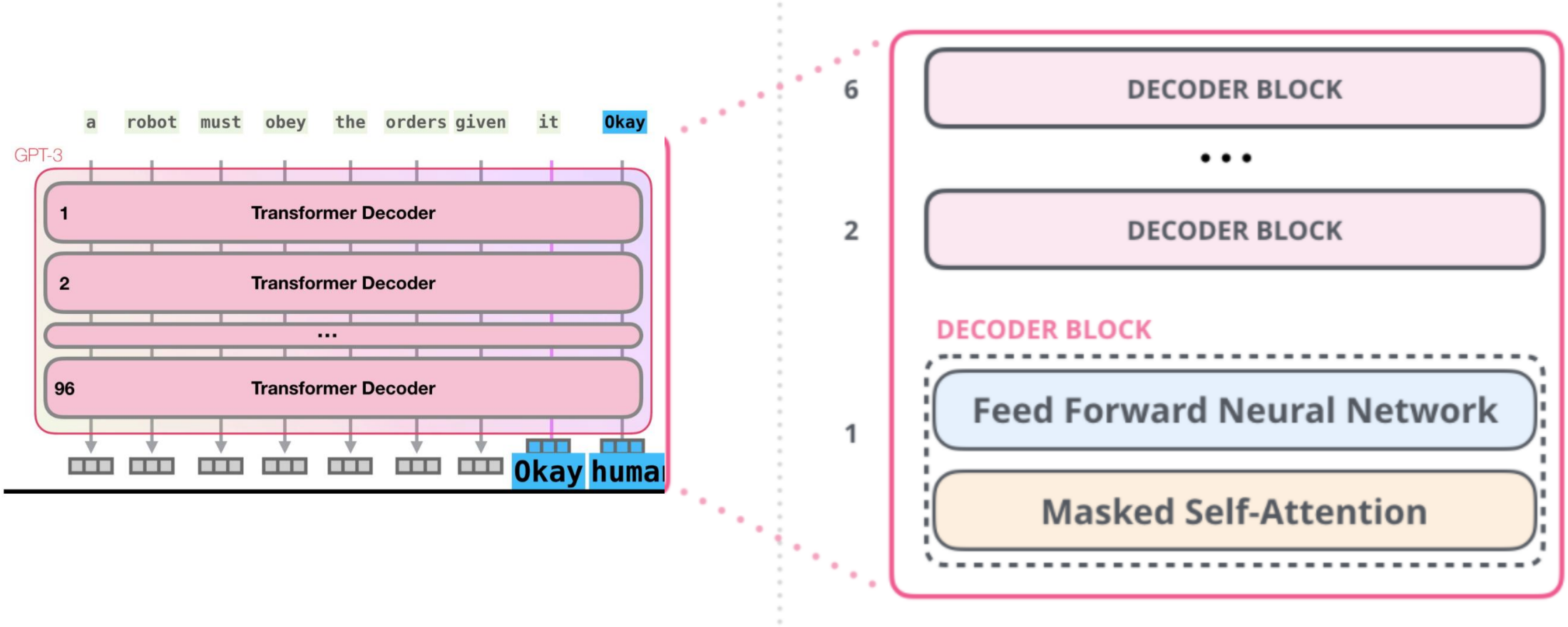
- Training is the process of exposing the model to lots of text.(300billion; 300,000,000,000, 3000억; tokens of text) That process has been completed. All the experiments you see now are from that **one trained model.**

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

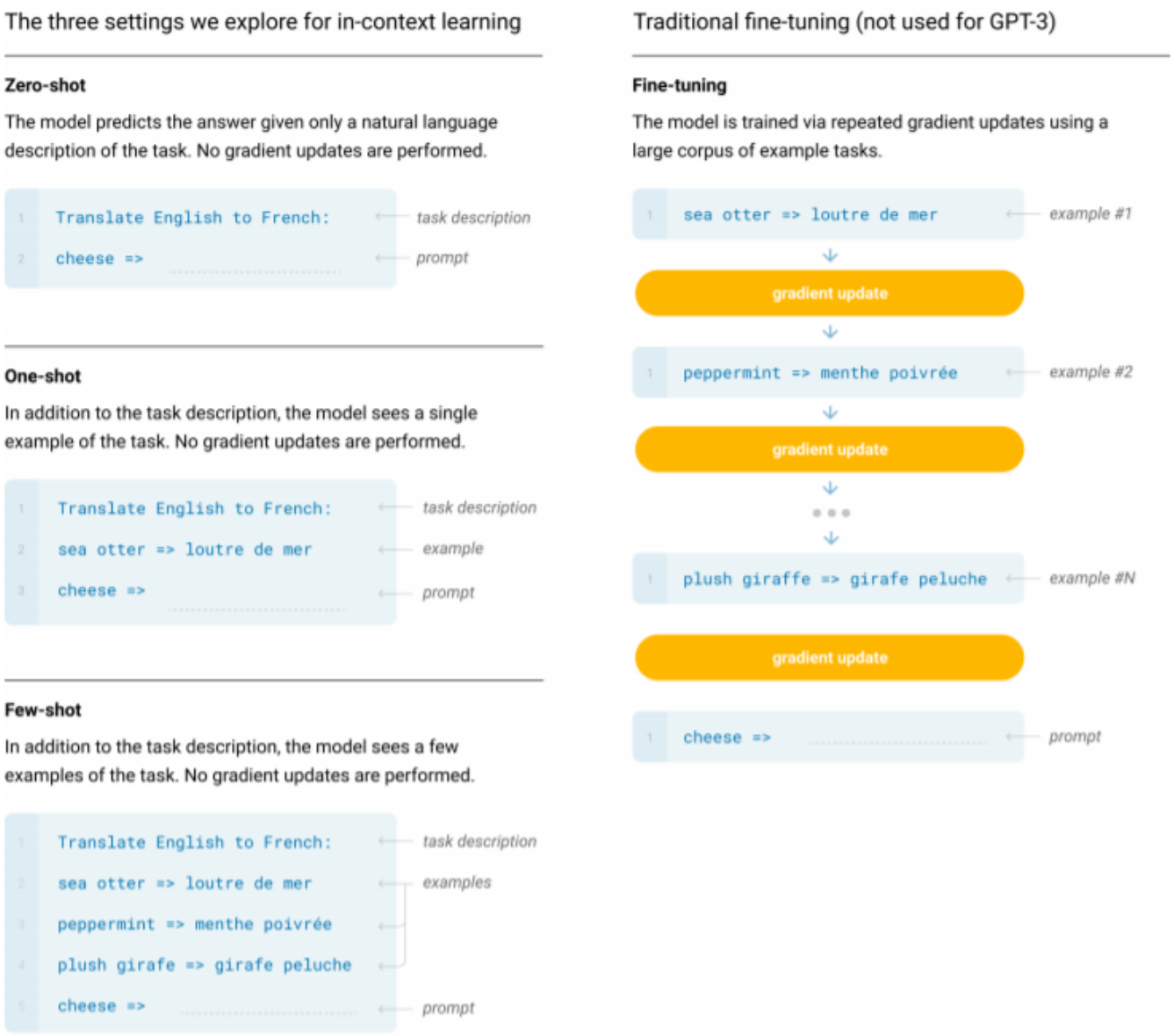
Table 2.2: Datasets used to train GPT-3. "Weight in training mix" refers to the fraction of examples during training that are drawn from a given dataset, which we intentionally do not make proportional to the size of the dataset. As a result, when we train for 300 billion tokens, some datasets are seen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

Architecture

- GPT-3's architecture is as same as GPT-2 (Transformer Decoder).
  The difference with GPT3 is the alternating dense and [sparse self-attention layers](). (번갈아 사용)

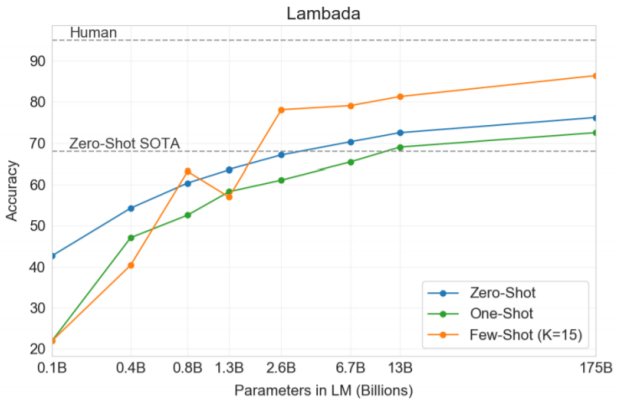


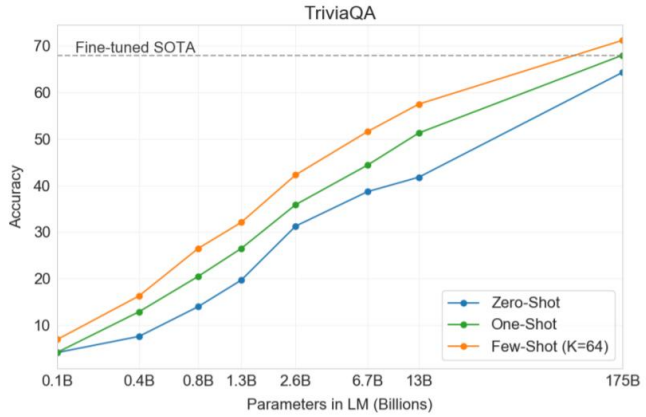
Zero, One, Few-shot

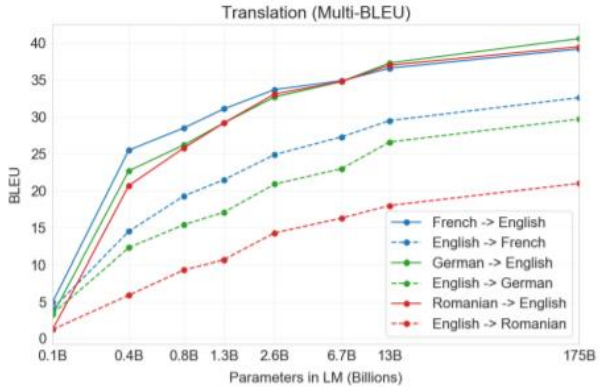


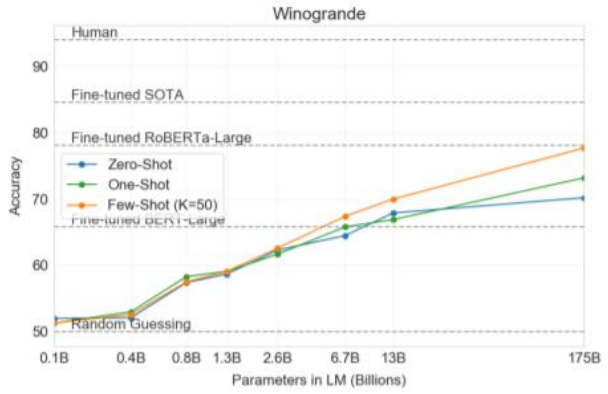
- Zero-shot, One-shot, Few-shot learning rather than Fine-tuning

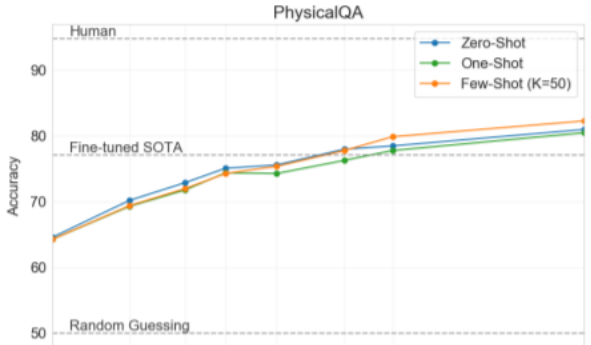
Task results











- Lambada (문장완성), HellaSwag (끝맺음 문장), StoryCloze(5문장중 끝맺기 문장)
- QA task (NaturalQS, WebQS, TriviaQA)
- Translation
- Winogrande (대명사 지칭)
- Common sense Reasoning (Physical QA)
- 기계 독해 (CoDA, DROP, QuAC, SQuADv2, RACE-h, RACE-m)
- Summarization (SuperGLUE)
- NLI (ANLI)
- Arithmetic (addition, subtraction, multiplication)
- Word Scrambling and manipulation task
- SAT analogies (동의어)
- News Article Generation
- 문법 교정

Impression

- Not Reasoning, 지식을 창조하진 않는다.
- 엄청나게 큰 DB, 검색엔진 (Google)대신
- Stored tr.data and Interpolation

Title: United Methodists Agree to Historic Split
Subtitle: Those who oppose gay marriage will form their own denomination
Article: After two days of intense debate, the United Methodist Church has agreed to a historic split - one that is expected to end in the creation of a new denomination, one that will be "theologically and socially conservative," according to The Washington Post. The majority of delegates attending the church's annual General Conference in May voted to strengthen a ban on the ordination of LGBTQ clergy and to write new rules that will "discipline" clergy who officiate at same-sex weddings. But those who opposed these measures have a new plan: They say they will form a separate denomination by 2020, calling their church the Christian Methodist denomination.
The Post notes that the denomination, which claims 12.5 million members, was in the early 20th century the "largest Protestant denomination in the U.S.," but that it has been shrinking in recent decades. The new split will be the second in the church's history. The first occurred in 1968, when roughly 10 percent of the denomination left to form the Evangelical United Brethren Church. The Post notes that the proposed split "comes at a critical time for the church, which has been losing members for years," which has been "pushed toward the brink of a schism over the role of LGBTQ people in the church." Gay marriage is not the only issue that has divided the church. In 2016, the denomination was split over ordination of transgender clergy, with the North Pacific regional conference voting to ban them from serving as clergy, and the South Pacific regional conference voting to allow them.

Figure 3.14: The GPT-3 generated news article that humans had the greatest difficulty distinguishing from a human written article (accuracy: 12%).

Google

voted to strengthen a ban on the ordination

The New York Times

United Methodists Tighten Ban on Same-Sex Marriage and Gay Clergy

By Timothy Williams and Elizabeth Dias

Feb. 26, 2019

714

ST. LOUIS — The United Methodist Church on Tuesday voted to strengthen its ban on gay and lesbian clergy and same-sex marriages, a decision that could split the nation's second-largest Protestant church.

Books

Zu meiner Mediathek hinzufügen | Rezension schreiben | Inhalt

E-BOOK KAUFEN – 3,40 CHF

Ergebnis 1 von 1 in diesem Buch für voted to strengthen a ban on the ordination of

Nach Druckexemplar suchen

0 Rezensionen
Rezension schreiben

The Killing of the Christian Church in America
von Gene Jackson

voted to strengthen a | Suche

Über dieses Buch

Meine Mediathek

Mein Verlauf

over the issue of same sex marriages and the ordaining of LGBTQ clergy.

At a recent February 2019 conference in St. Louis, UMC officials and lay members voted to strengthen prohibitions and to ban LGBTQ people from being ordained and ministers from performing same-sex weddings within the church.

For several years, the LGBTQ movement and their supporters within the church have been pushing to allow persons of any sexual orientation to be married in the church and for gays to participate in leadership roles.

In some areas, churches are already in violation of their Book of Discipline teachings by the appointing of gay ministers and the

Impression

Context →	Q: What is 98 plus 45? A:
Target Completion →	143

Figure G.44: Formatted dataset example for Arithmetic 2D+

Context →	Q: What is 95 times 45? A:
Target Completion →	4275

Figure G.45: Formatted dataset example for Arithmetic 2Dx

Context →	Q: What is 509 minus 488? A:
Target Completion →	21

Figure G.46: Formatted dataset example for Arithmetic 3D-

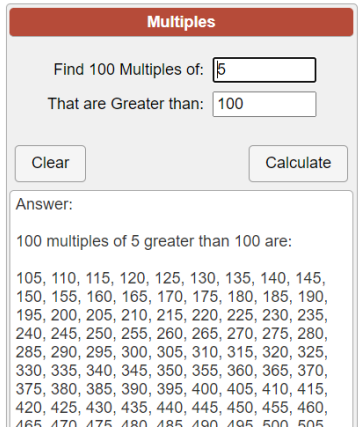
Context →	Q: What is 556 plus 497? A:
Target Completion →	1053

98 45 95 45 50 12

Google

98 45 95 45 50 12

Multiples Calculator



Calculator Use

The multiples of numbers calculator will find 100 multiples of a positive integer. For example, the multiples of 3 are calculated 3x1, 3x2, 3x3, 3x4, 3x5, etc., which equal 3, 6, 9, 12, 15, etc. You can designate a minimum value to generate multiples greater than a number. For example, to find 100 multiples of 36 that are greater than 1000 you will get: 1008, 1044, 1080, 1116, 1152, 1188, 1224, 1260, 1296, 1332, 1368, 1404, etc.

Here is a list of the first 20 multiples of the integers 1 through 20.

Multiples of 1: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20

Multiples of 2: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40

Multiples of 3: 3, 6, 9, 12, 15, 18, 21, 24, 27, 30, 33, 36, 39, 42, 45, 48, 51, 54, 57, 60

Introduction

- 첫 째, 지금과 같은 방식에서는 새 태스크를 풀 때마다 많은 **라벨링된 데이터가** 필요하다.
- 두 번째, 사전학습 후 **fine-tuning 하는** 방법에서는 접근법에서 모델은 사전학습 중 대량의 지식을 흡수하지만, 아주 작은 태스크 분포에 대해 fine-tune 된다. 그 결과, 모델이 크다고 해서 out-of-distribution 문제를 더 잘 일반화하지 못하다는 연구 결과도 있다. 이는 훈련 데이터의 분포에 대해 한정지어진 모델이 그 외의 영역은 잘 일반화하지 못한다는 증거이고, 벤치마크 태스크에 대해서는 마치 성적이 좋은 것처럼 보일지언정 사람이 볼 때는 그 성능이 과장된 것처럼 느낄 수 있다.

https://littlefoxdiary.tistory.com/44

Introduction

- 세 번째, 사람은 대부분의 언어 태스크를 하기 위해 '예제 데이터'를 많이 필요로 하지 않는다. 간단한 지시사항, 이를테면 "이 문장이 긍정적인 감성을 담고 있는지, 부정적인 감성을 담고 있는지 말해보세요"과 같은 문장만으로도 태스크를 어느 정도 잘 해낼 수 있다. 뿐만 아니라 대화를 하다가 간단한 덧셈을 실시하는 등, 많은 태스크와 기술을 왔다 갔다 하며 사용할 수 있다. 언어 모델이 유용하려면, NLP 시스템 역시 이러한 **유연성과 일반성을** 가져야 할 것이다.
- GPT-2에서는 이러한 방법을 "**in-context learning**" 방식으로 진행했는데, 사전학습 모델에 풀고자 하는 태스크를 텍스트 인풋으로 넣는 방식이다. 하지만 안타깝게도 결과들은 몇몇 태스크에서는 fine-tuning 접근법에 미치지 못하는 아쉬운 성능을 보였다.

https://littlefoxdiary.tistory.com/44

Introduction

- 최근 NLP 연구의 또 다른 트렌드는 모델 크기를 키우는 것이다. 트랜스 포머를 이용하면 모델 사이즈를 크게 늘릴 수 있고 파라메터 수가 1억 개(GPT-1), 3억 개(BERT), 14억 개(GPT-2), 80억 개(Megatron), 110억 개(T5), 170억 개(Project Turing)까지 늘어나며 다운스트림 태스크에서의 성능은 점점 더 좋아졌다. in-context learning 방식은 최대한 다양한 스킬과 태스크를 모델의 파라메터에 저장해야 하고, **모델의 스케일이 증가할 때 성능이 증가할 가능성이 있다.**
- GPT-3는 **1750억개 parameters**, Zero-shot (예제 없음, 테스크 정의만), One-shot (하나의 예제), Few-shot learning (in-context learning) (10-100개정도의 예제)

https://littlefoxdiary.tistory.com/44

Approach

1) Fine-Tuning(FT) - NLP에서 가장 보편적인 방법으로, 사전 학습된 모델의 웨이트를 다운스트림 태스크에 대해 미세 조정하는 것. 보통 수천 개의 라벨링 된 데이터를 사용한다. 이 방법은 성능 향상에 크게 도움되지만, 매 태스크마다 라벨링된 데이터가 지나치게 많이 필요하다는 단점이 있다.

2) Few-Shot(FS) - 모델은 예시 태스크를 보게 되지만, 가중치 업데이트는 일어나지 않는다. K(10~100) 개의 예제를 context (2048 토큰까지 처리) 부분에 주고, 추론하려는 example의 결과를 완성하도록 하는 접근법이다. 이 방법에서는 태스크에 대한 소량의 예제만이 필요하다. 하지만, 대부분의 모델에서 few-shot 성능은 fine-tuning 결과를 따라가지 못한다.
FS 세팅에서 모델에게 번역 태스크를 시키고자 한다면 context 부분에는 다음과 같은 입력을 넣는다.

"한국어를 영어로 번역하라: 집에 가고 싶어 -> I want to go home. 배고파 -> I am hungry 치킨 사줘 -> _____ "

모델은 문맥 인풋에 있는 예시들을 보고 _____ 부분에 "Buy me fried chicken"을 채워 넣어야 한다.

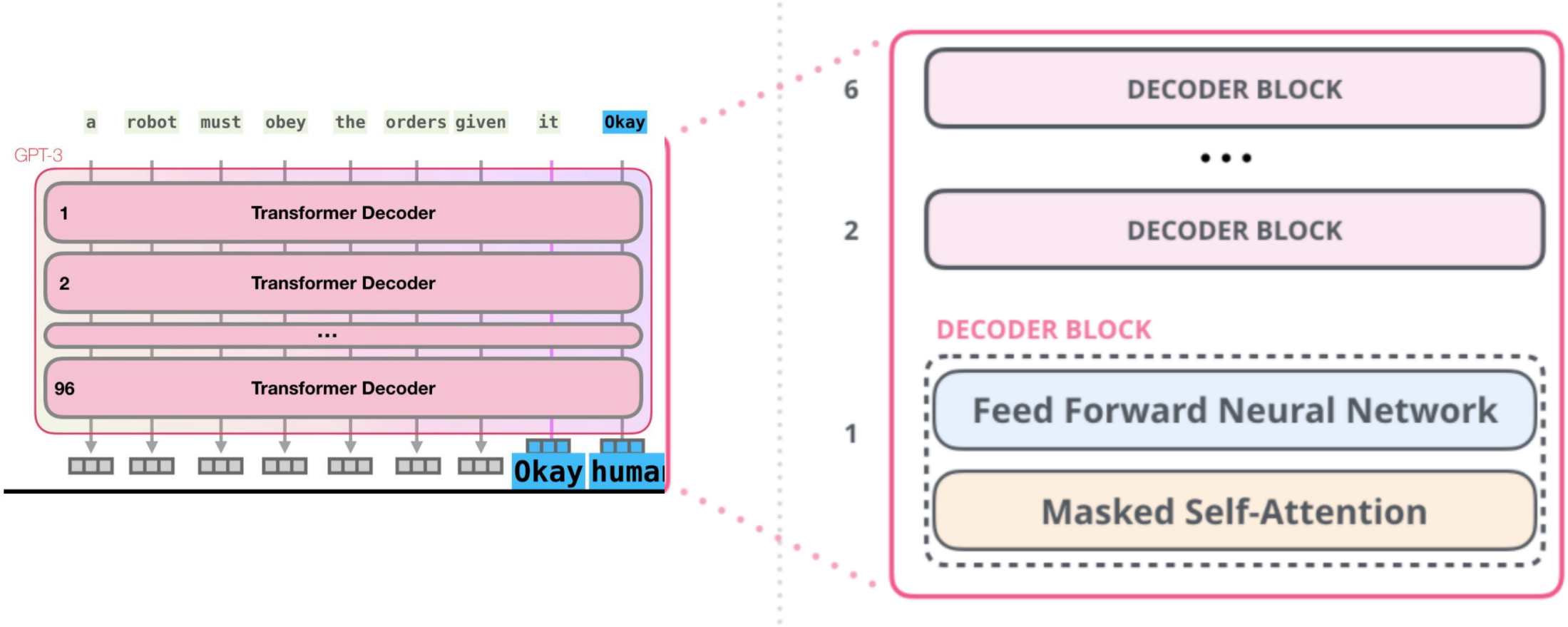
3) One-Shot(1S) - FS 세팅과 같으나, 하나의 예제만을 예시로 준다.

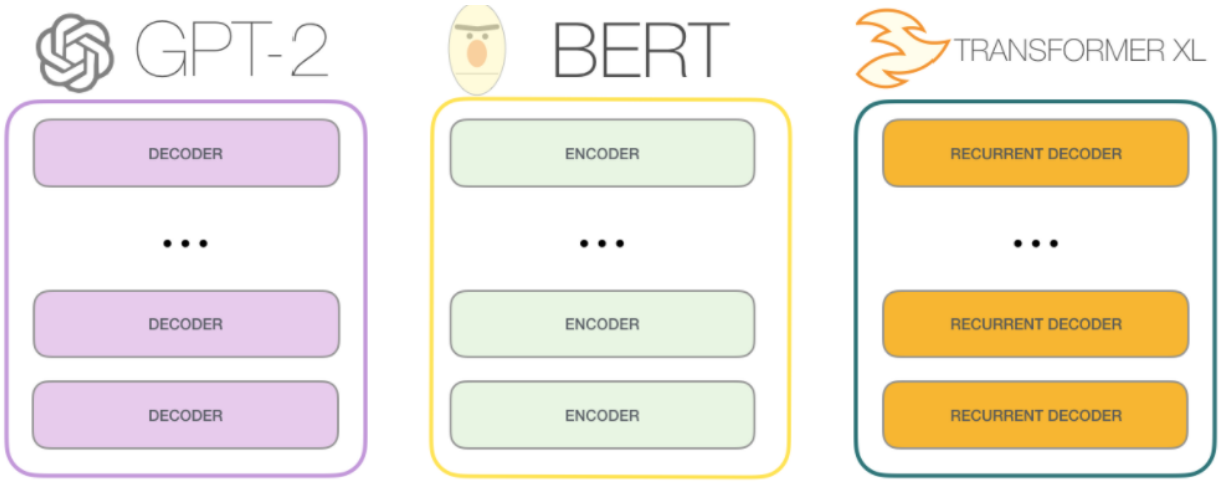
4) Zero-Shot(0S) - 태스크에 대한 예시는 주지 않고, 태스크를 설명하는 자연어 문구만을 준다. 이 방법은 엄청나게 편리할 뿐만 아니라 잠재적으로 강건하고 사전학습 데이터에 편재할 수 있는 좋지 않은 상관관계를 피하게 한다. 하지만, 이는 굉장히 어려운 과제이고 아마 몇몇 과제는 사람조차도 지시사항만으로 푸는 데에 어려움을 느낄 수 있다.

https://littlefoxdiary.tistory.com/44

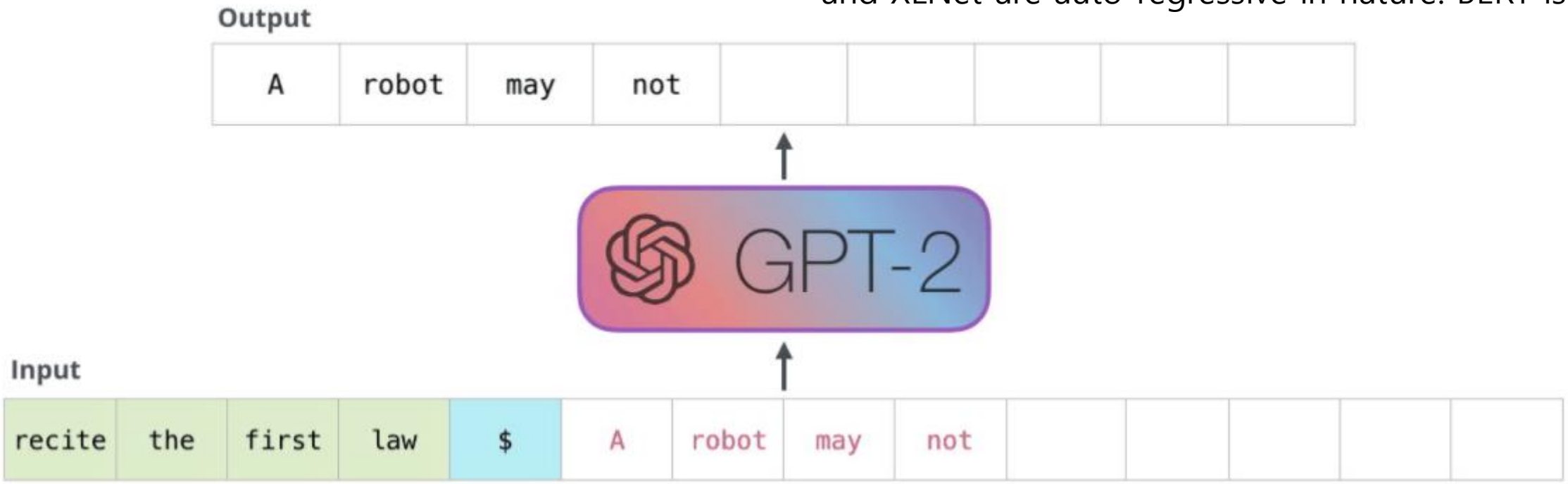
Architecture

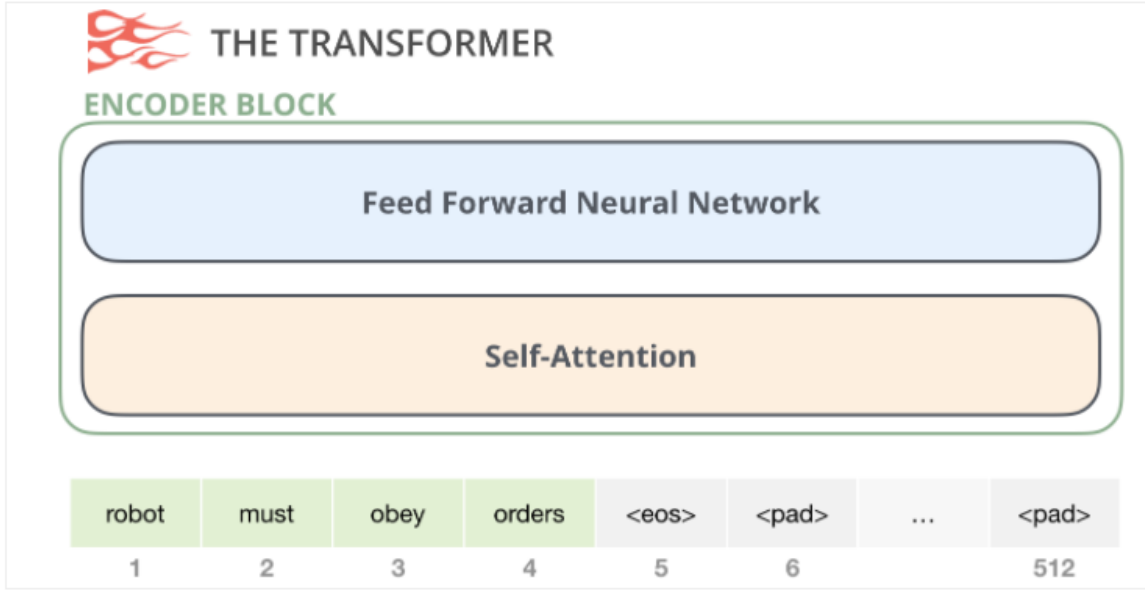
GPT-3's architecture is as same as GPT-2 (Transformer Decoder). The difference with GPT3 is the alternating dense and [sparse self-attention layers](). (번갈아 사용)



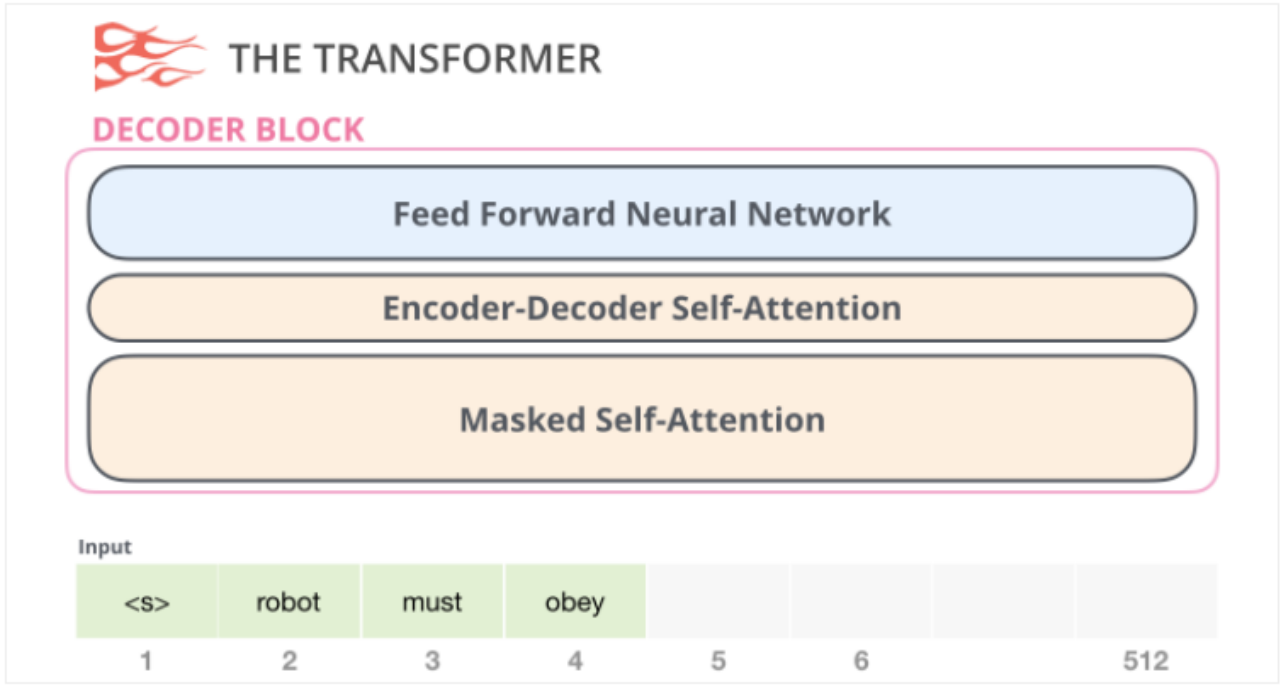


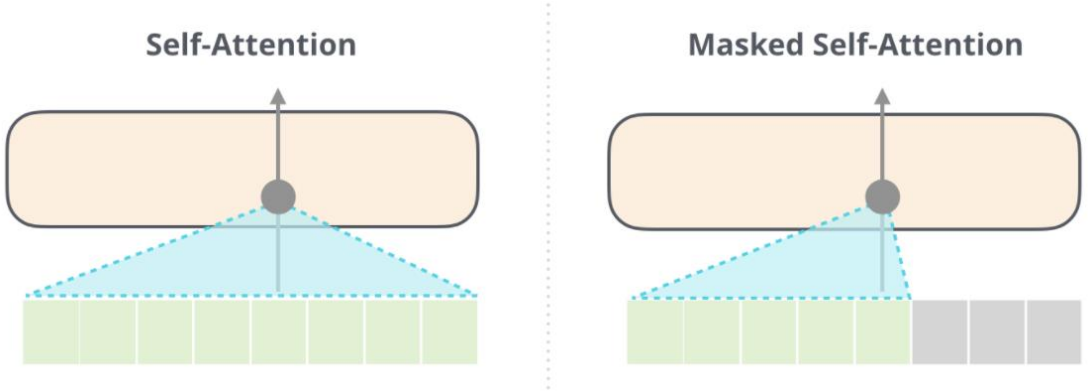
The way these models actually work is that after each token is **produced, that token is added** to the sequence of inputs. And that new sequence becomes the input to the model in its next step. This is an idea called "auto-regression".
The GPT2, and some later models like TransformerXL and XLNet are auto-regressive in nature. BERT is not.



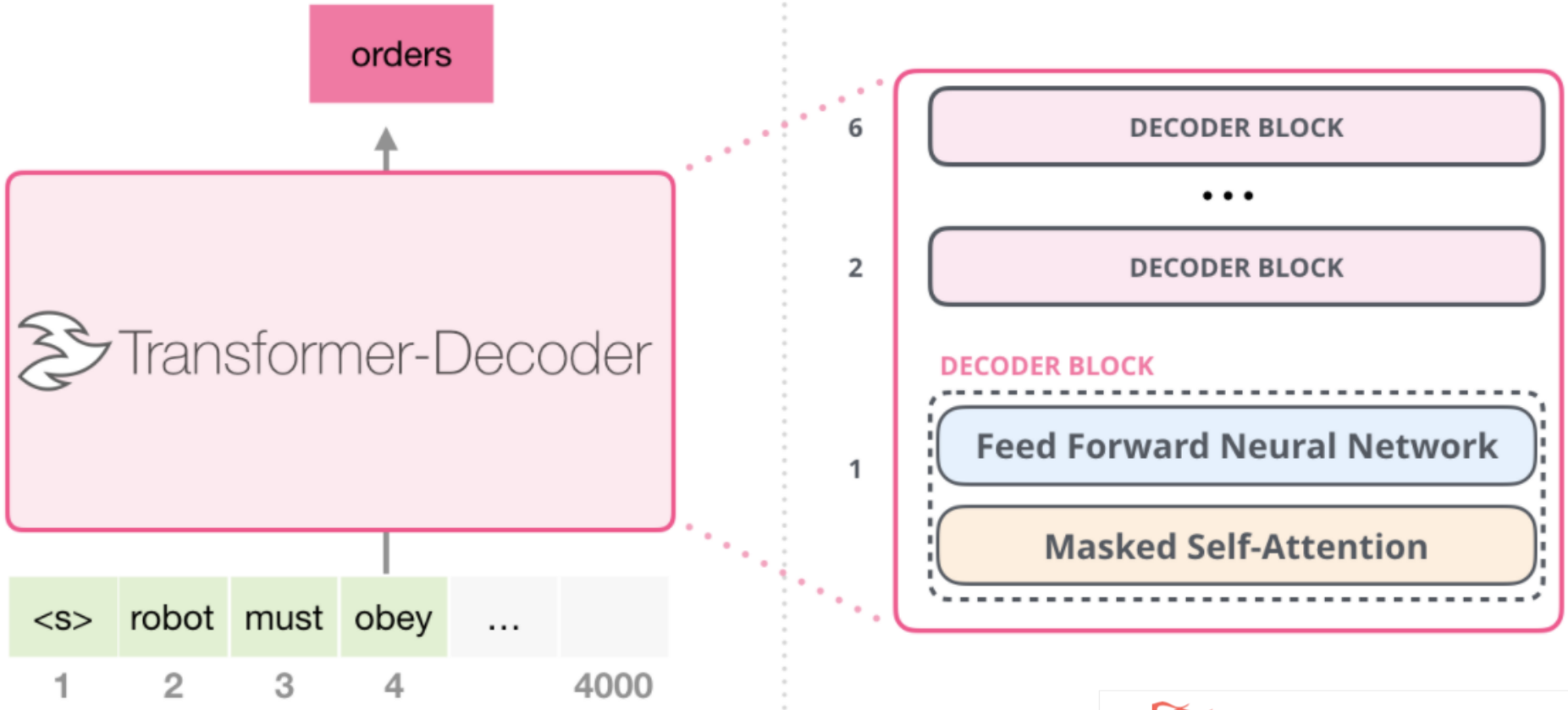


An encoder block from the original transformer paper can take inputs up until a certain max sequence length (e.g. 512 tokens). It's okay if an input sequence is shorter than th... limit, we can just pad the rest of the sequence.

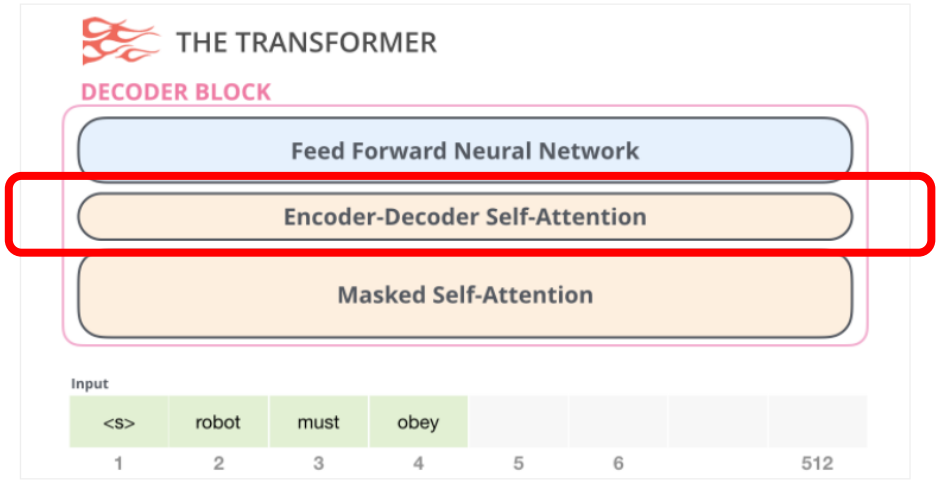


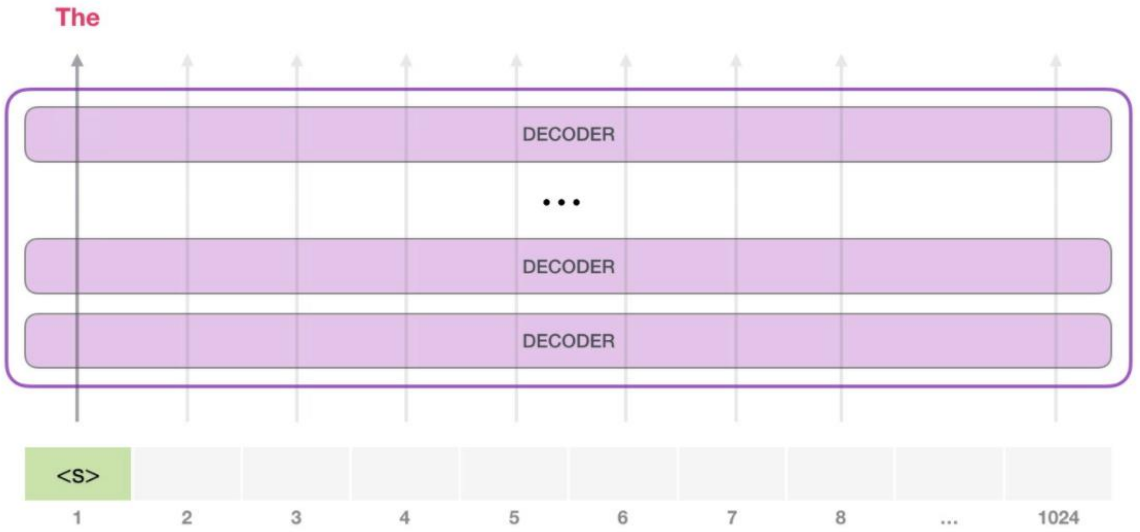


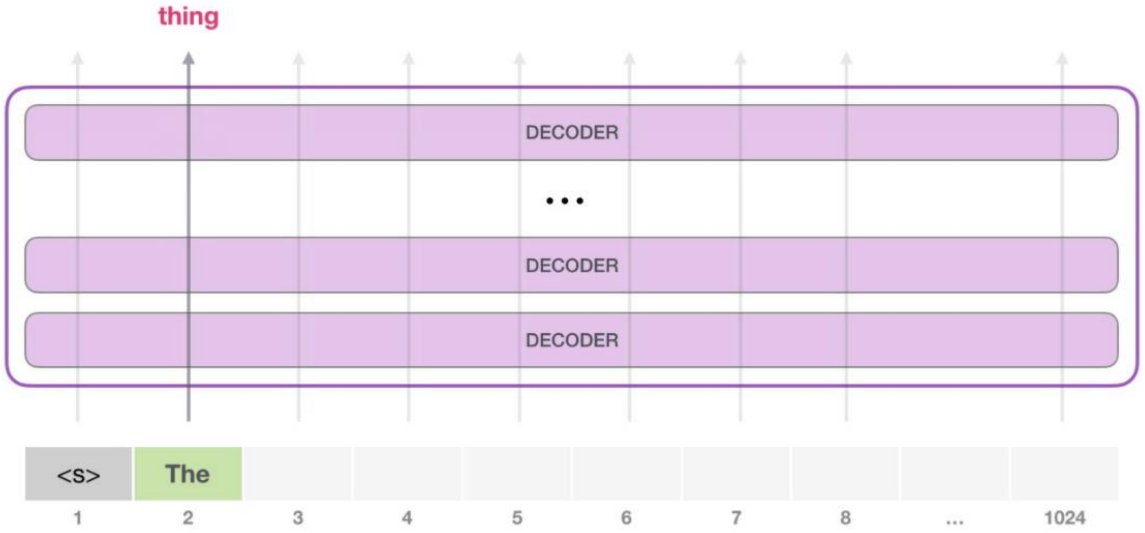
*Masked self-attention
not by changing the word to [mask] like BERT, **but** by interfering in the self-attention calculation **blocking information from tokens**



These blocks were very similar to the original decoder blocks, **except they did away with that second self-attention layer**. (Encoder-Decoder self-attention)







The GPT-2 can process **1024 tokens**. Each token flows through all the decoder blocks along its own path.
GPT3 (2048 tokens)

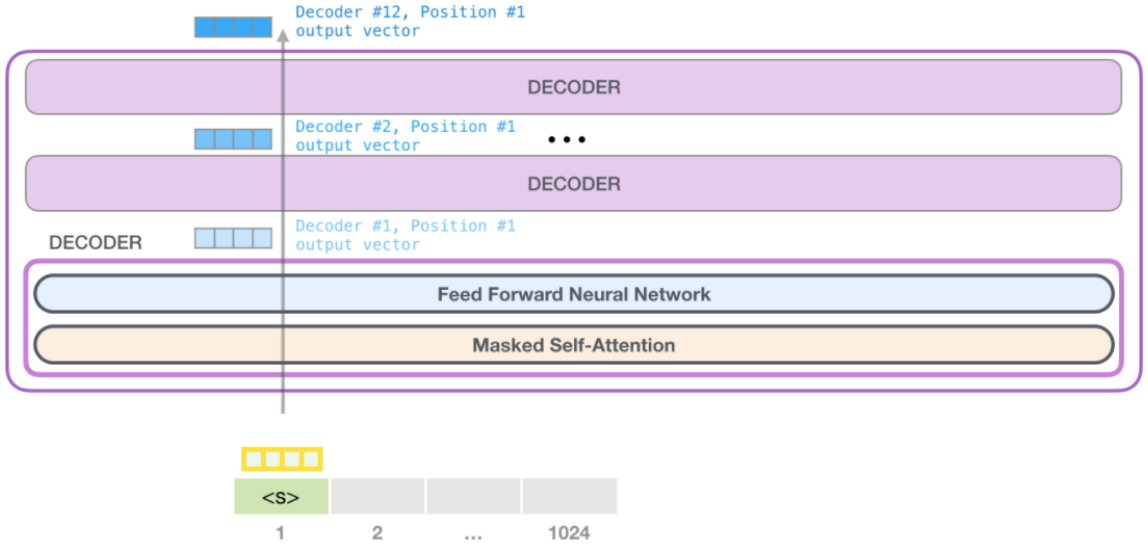
The model only has one input token, so that path would be the only active one.

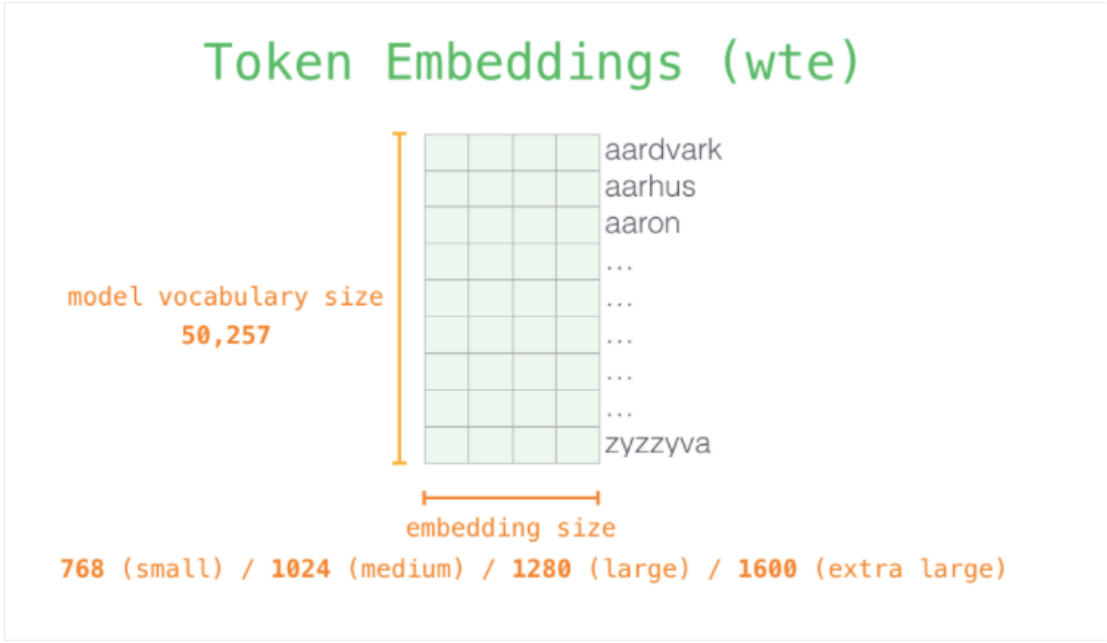
After process, it selected the token with the highest probability, ‘the’. GPT-2 has a parameter **called top-k** that we can use to have the model consider **sampling words other than the top word**

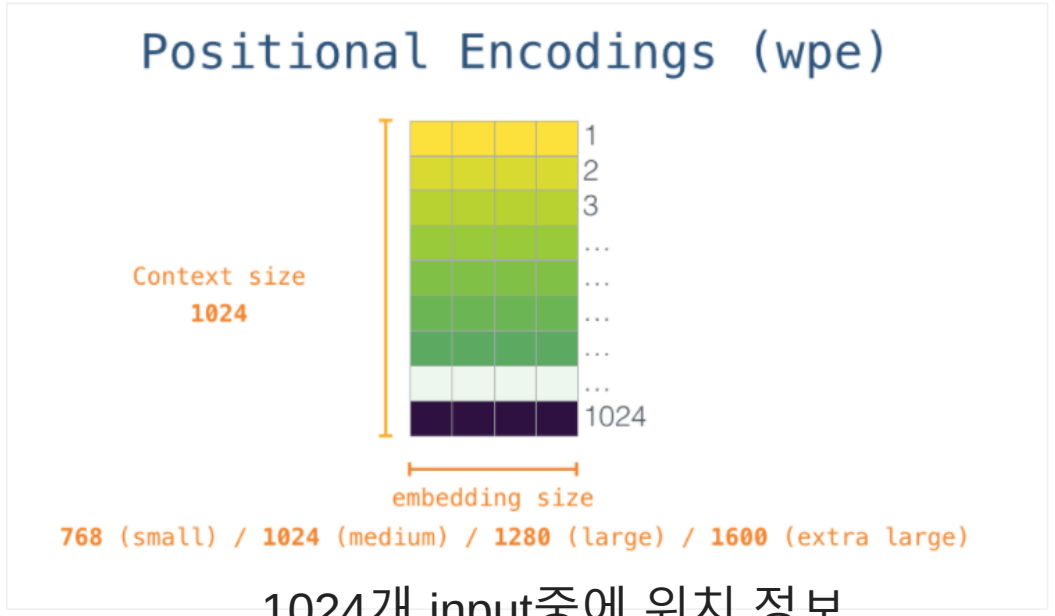
In the next step, **we add the output from the first step to our input sequence**, and have the model make its next prediction. “The” -> “thing”

Each layer of GPT-2 has **retained its own interpretation of the first token** and will use it in processing the second token (Self-attention)

Input encoding

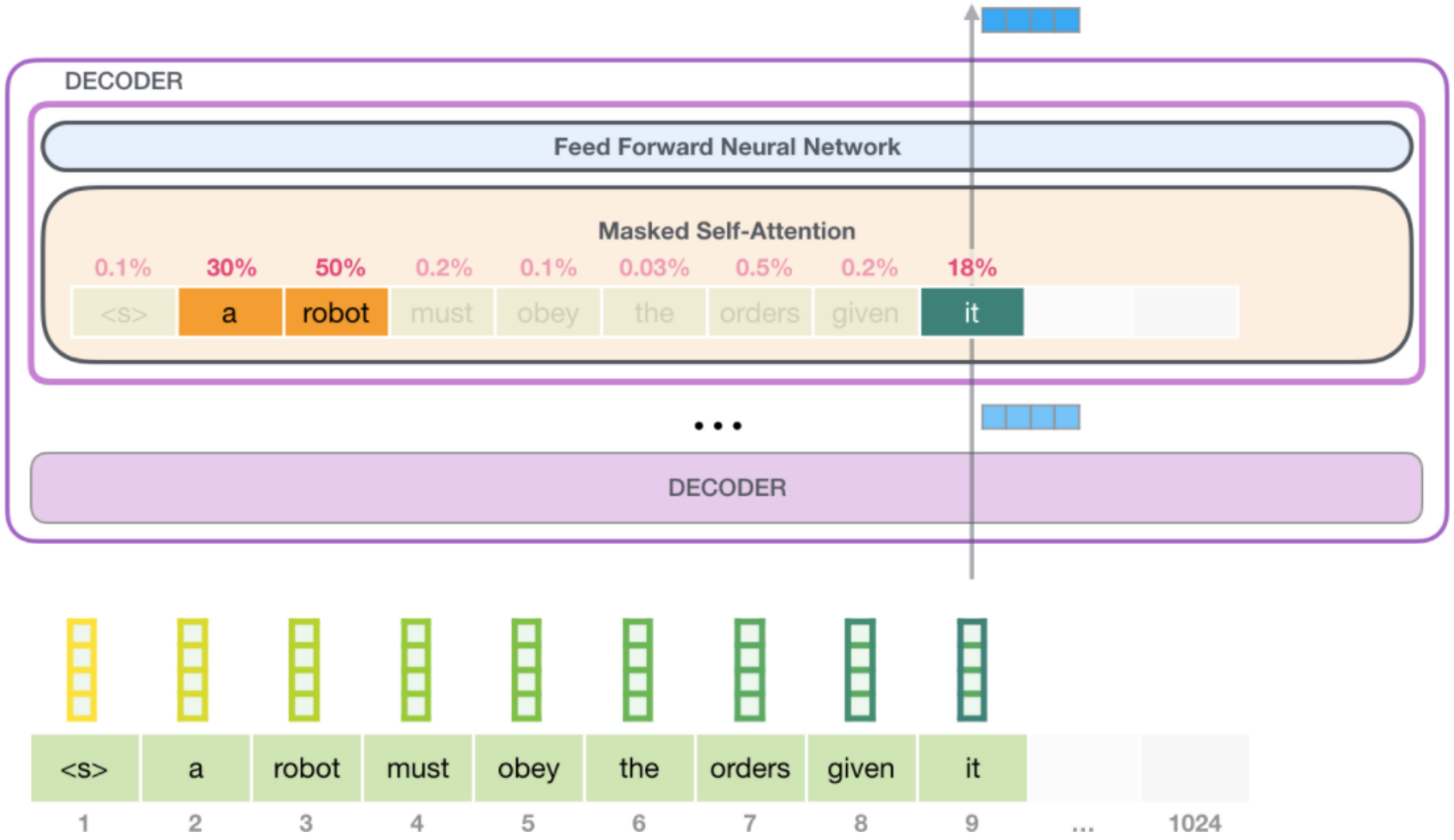




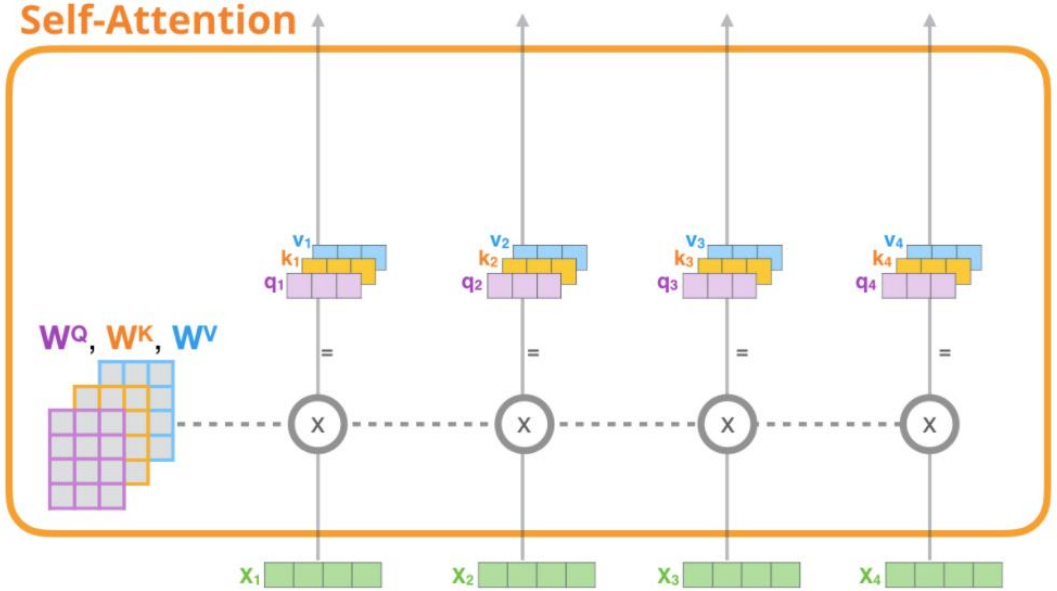


1024개 input중에 위치 정보

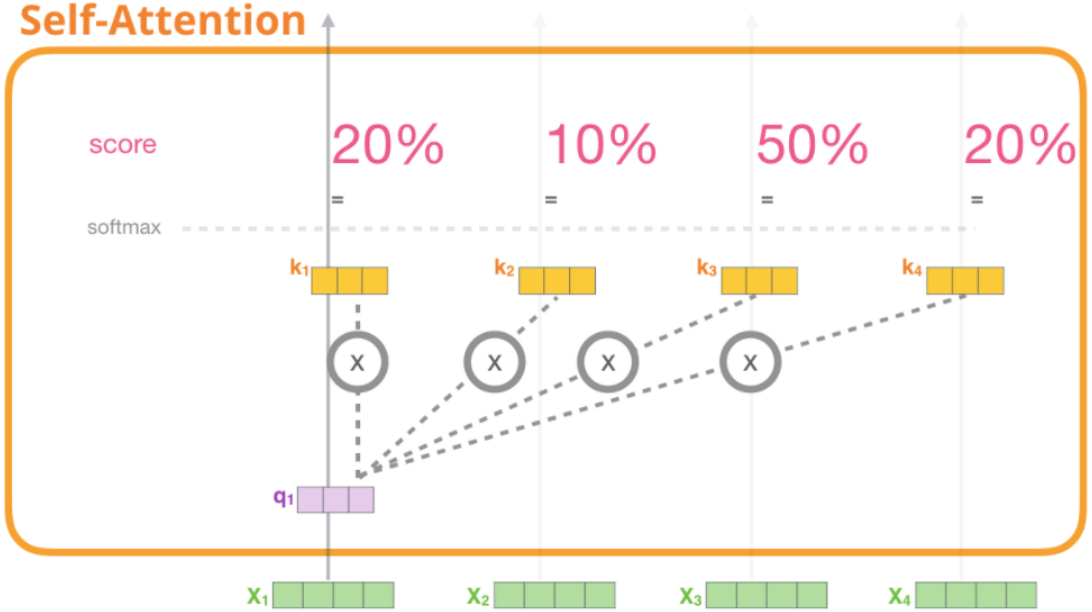
Self-attention

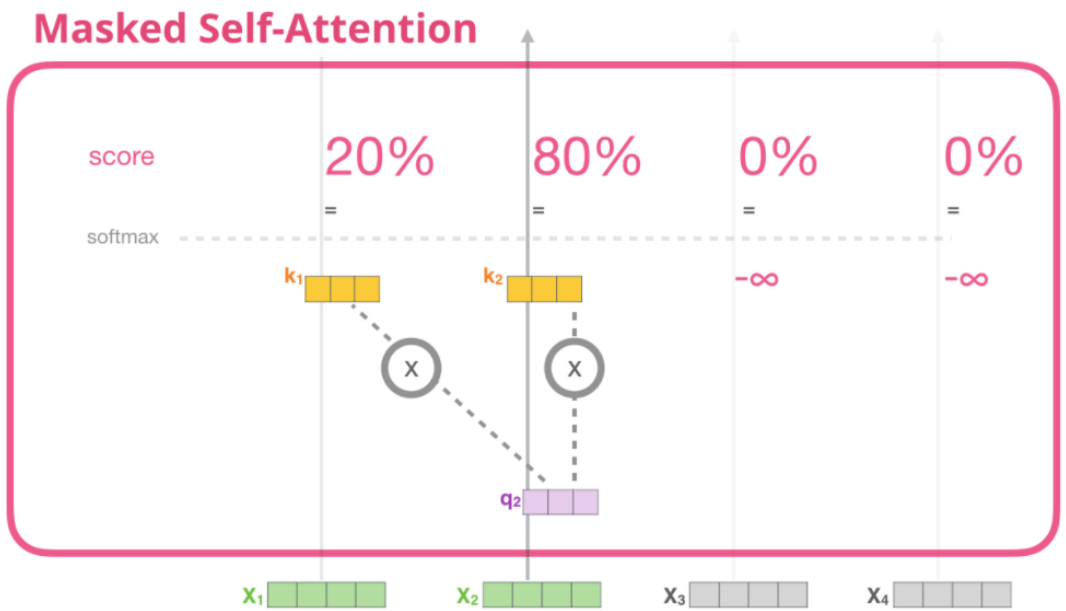


1) For each input token, create a query vector, a key vector, and a value vector by multiplying by weight Matrices W^Q, W^K, W^V

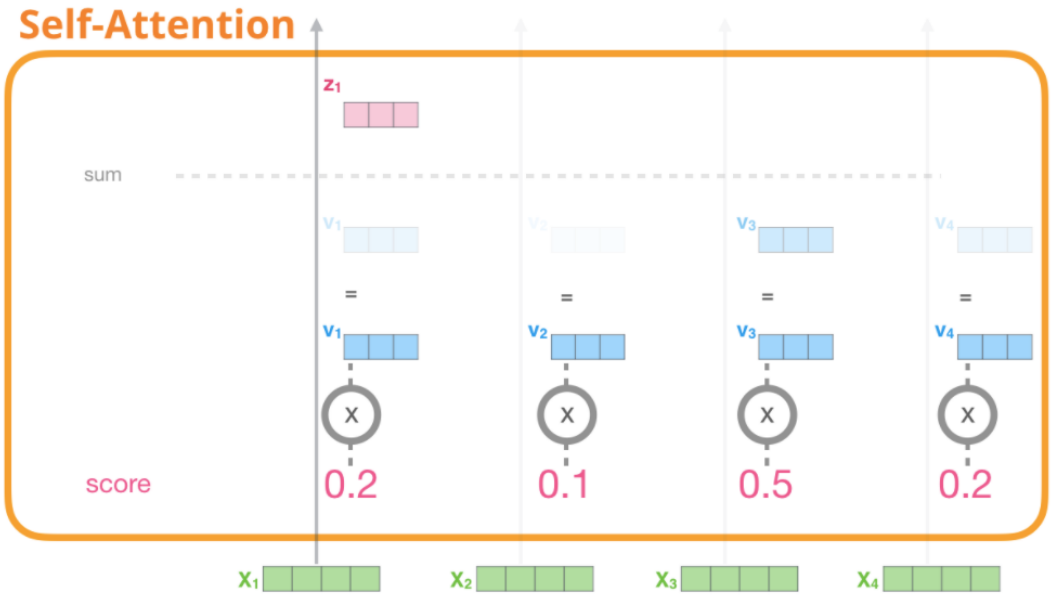


2) Multiply (dot product) the current query vector, by all the key vectors, to get a score of how well they match





3) Multiply the value vectors by the scores, then sum up



Queries

robot	must	obey	orders

X

Keys

robot	must	obey	orders
robot	must	obey	orders
robot	must	obey	orders
robot	must	obey	orders

=

Scores (before softmax)

0.11	0.00	0.81	0.79
0.19	0.50	0.30	0.48
0.53	0.98	0.95	0.14
0.81	0.86	0.38	0.90

Scores (before softmax)

0.11	0.00	0.81	0.79
0.19	0.50	0.30	0.48
0.53	0.98	0.95	0.14
0.81	0.86	0.38	0.90

Apply Attention Mask →

Masked Scores (before softmax)

0.11	-inf	-inf	-inf
0.19	0.50	-inf	-inf
0.53	0.98	0.95	-inf
0.81	0.86	0.38	0.90

Masked Scores (before softmax)

0.11	-inf	-inf	-inf
0.19	0.50	-inf	-inf
0.53	0.98	0.95	-inf
0.81	0.86	0.38	0.90

Softmax (along rows) →

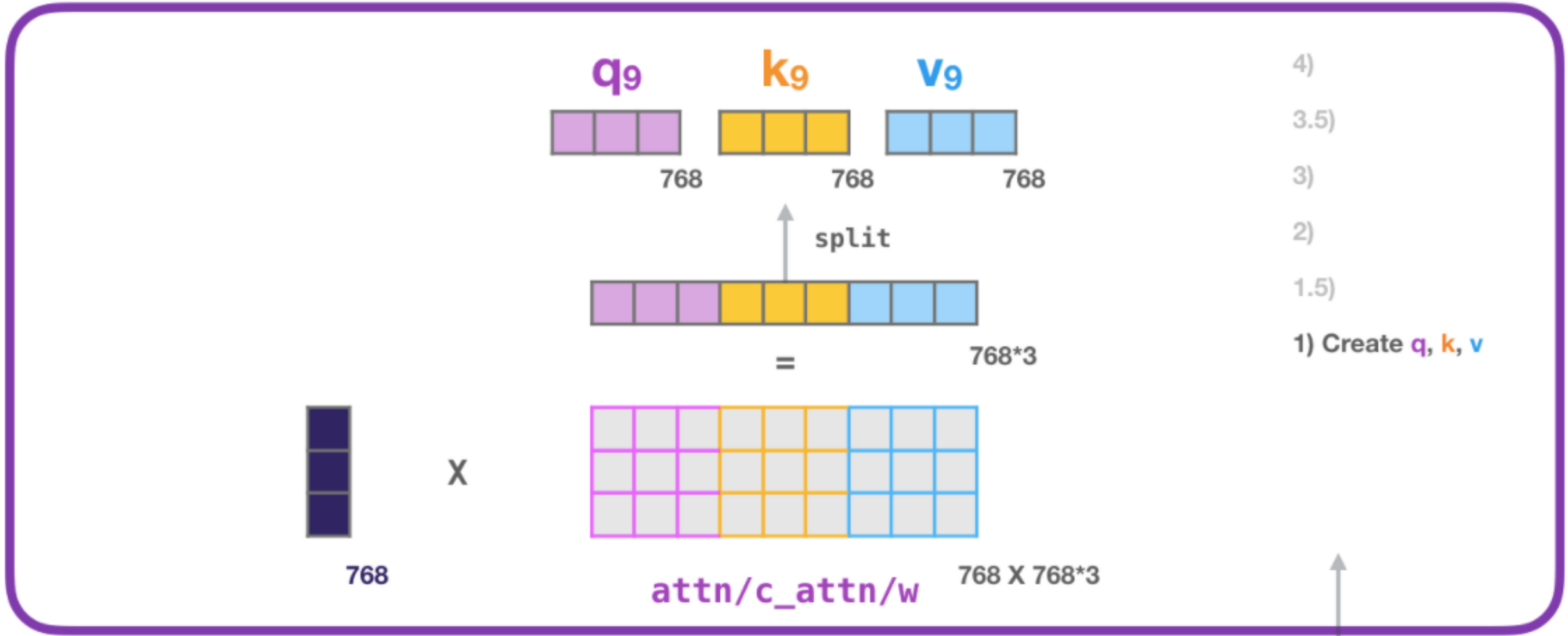
Scores

1	0	0	0
0.48	0.52	0	0
0.31	0.35	0.34	0
0.25	0.26	0.23	0.26

Score * Value -> Z

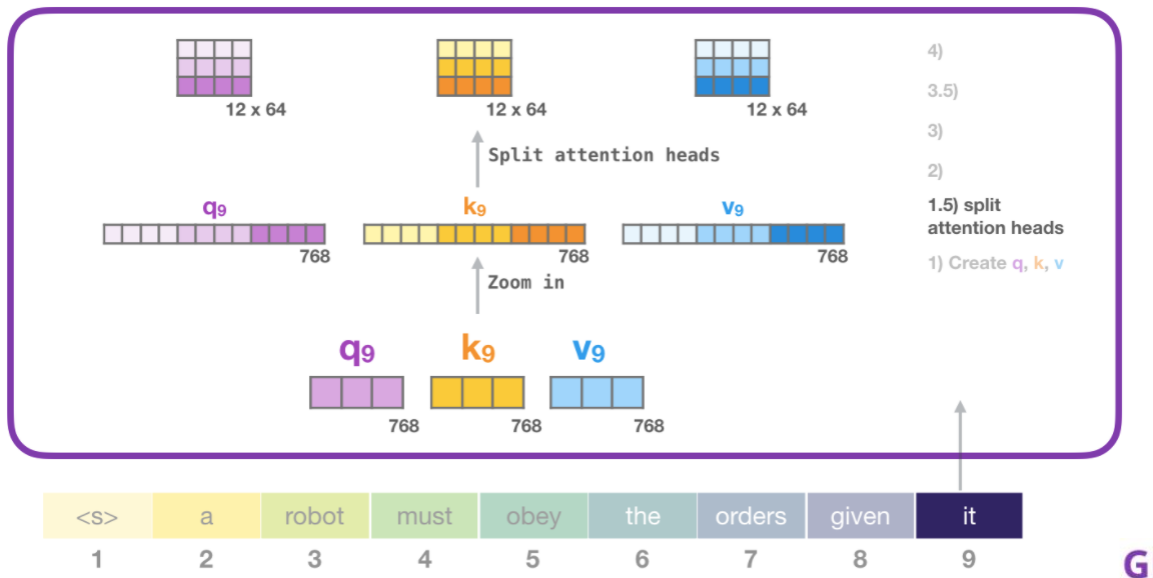
GPT2 Self-Attention

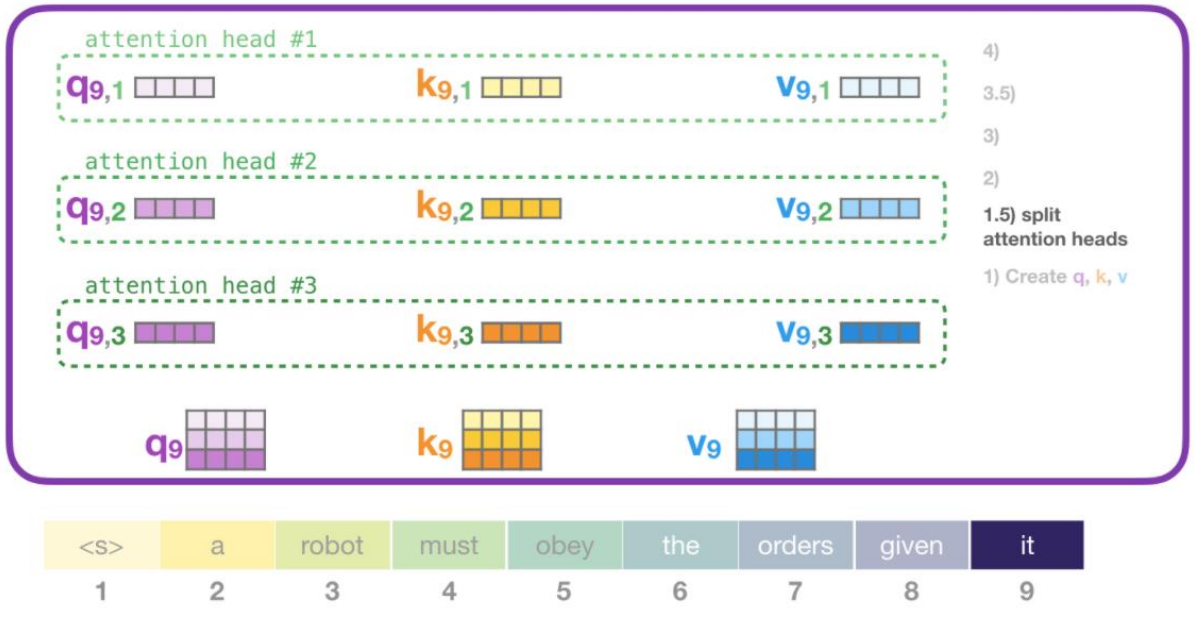
1) Create Q,K,V



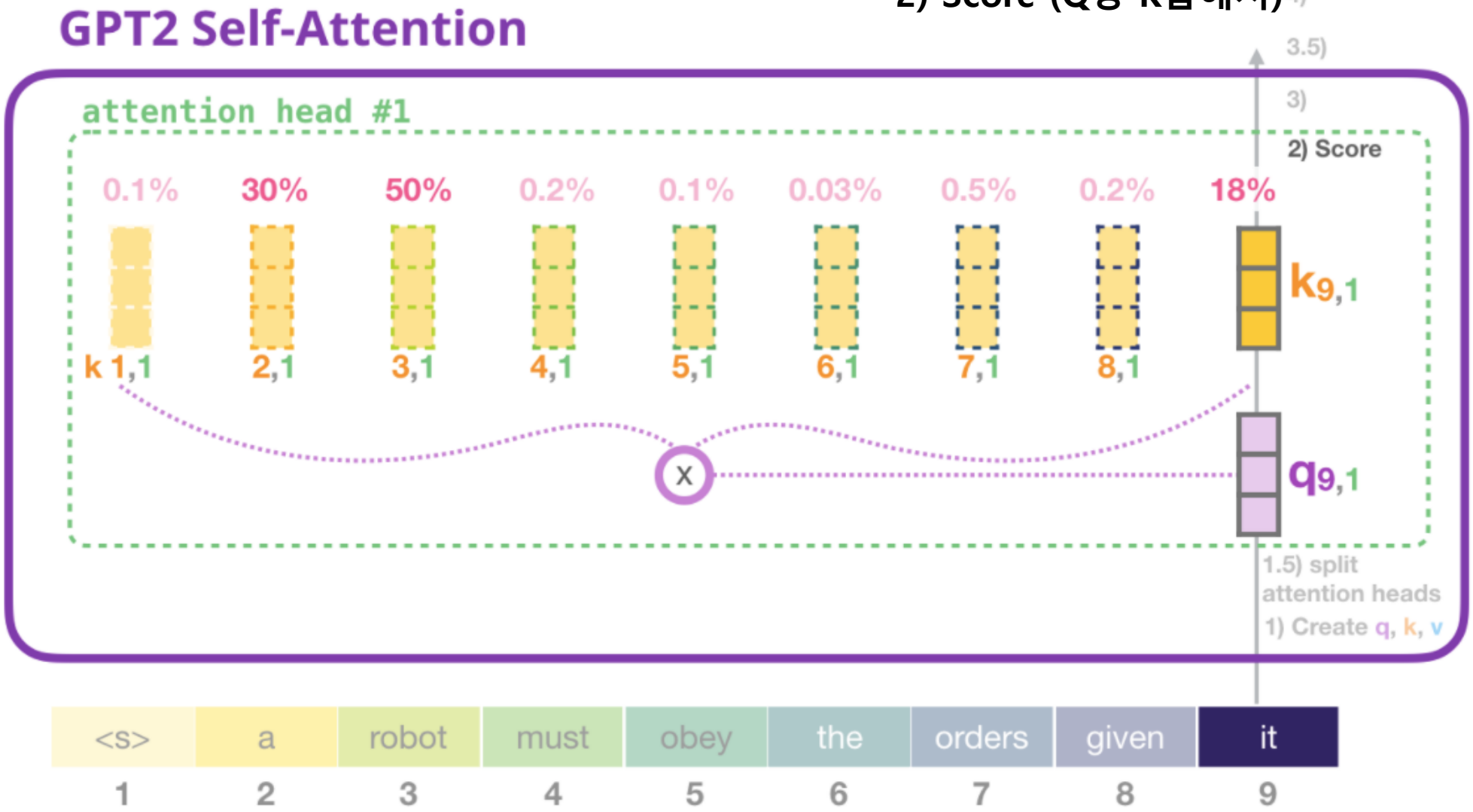
1) Create Q,K,V

1.5) Split attention head

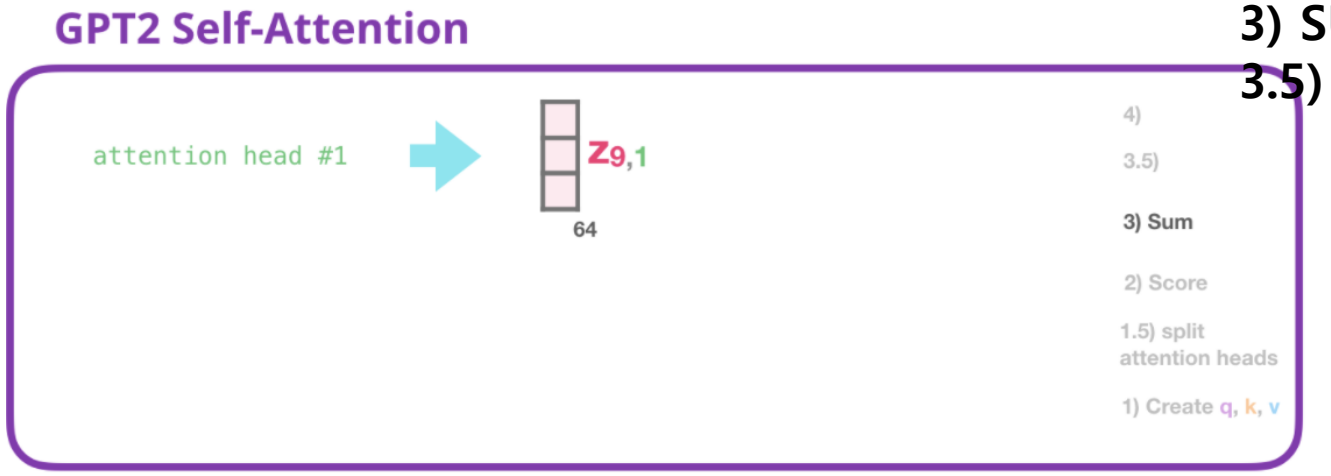


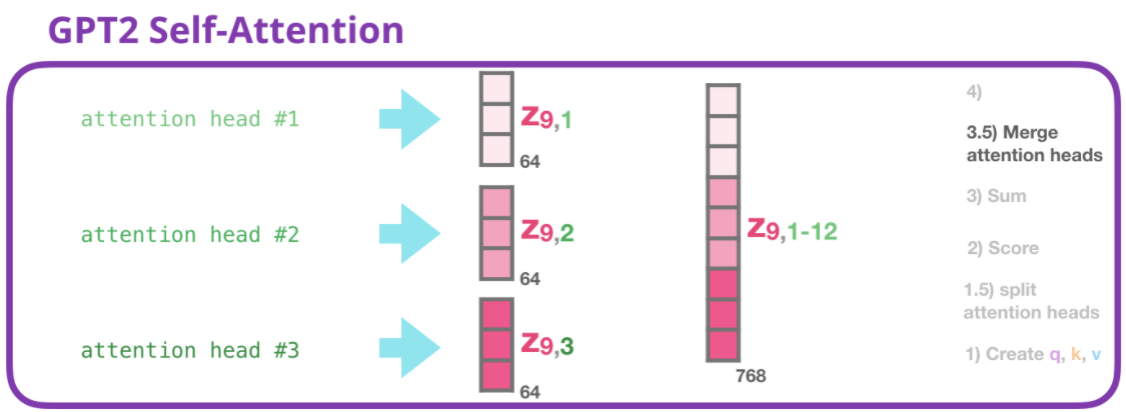


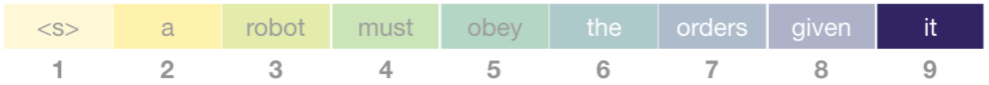
1) Create Q,K,V
1.5) Split attention head
2) Score (Q랑 K곱해서)

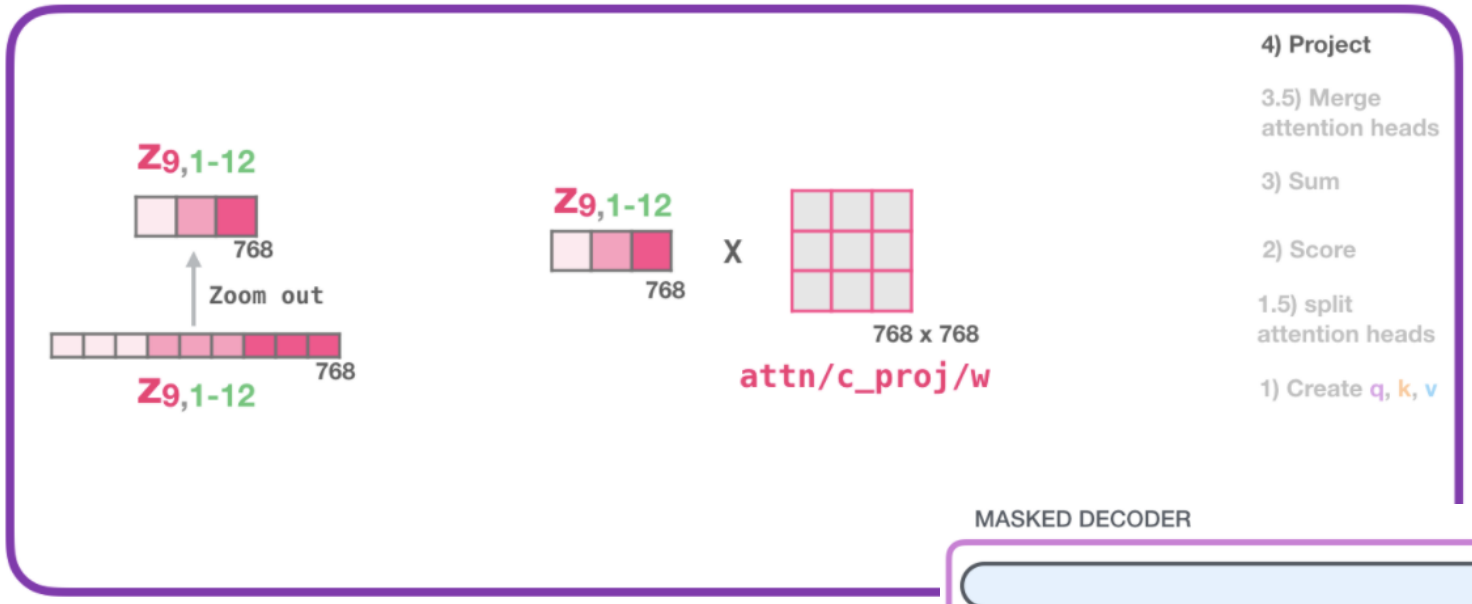


1) Create Q,K,V
1.5) Split attention head
2) Score (Q랑 K곱해서)
3) SUM (Z구함)
3.5) Merge attention heads

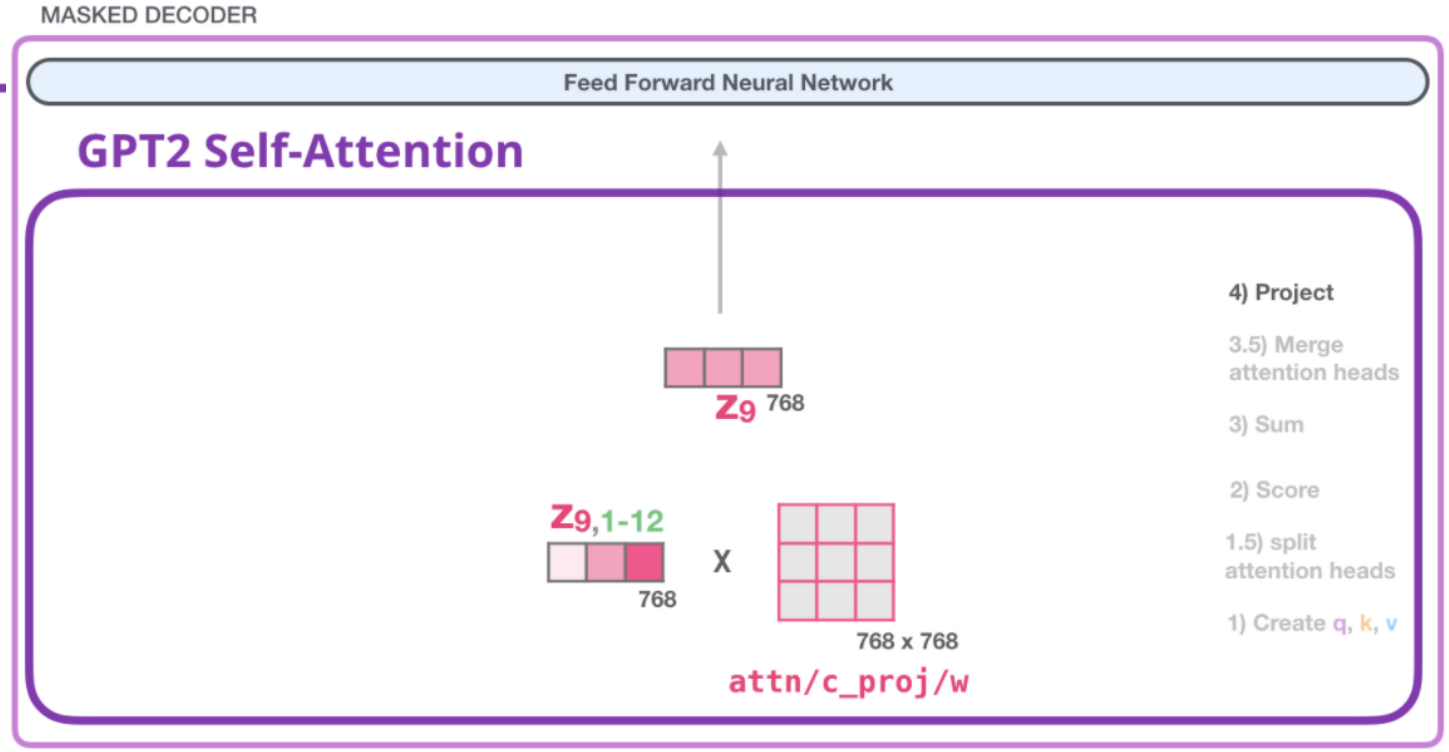




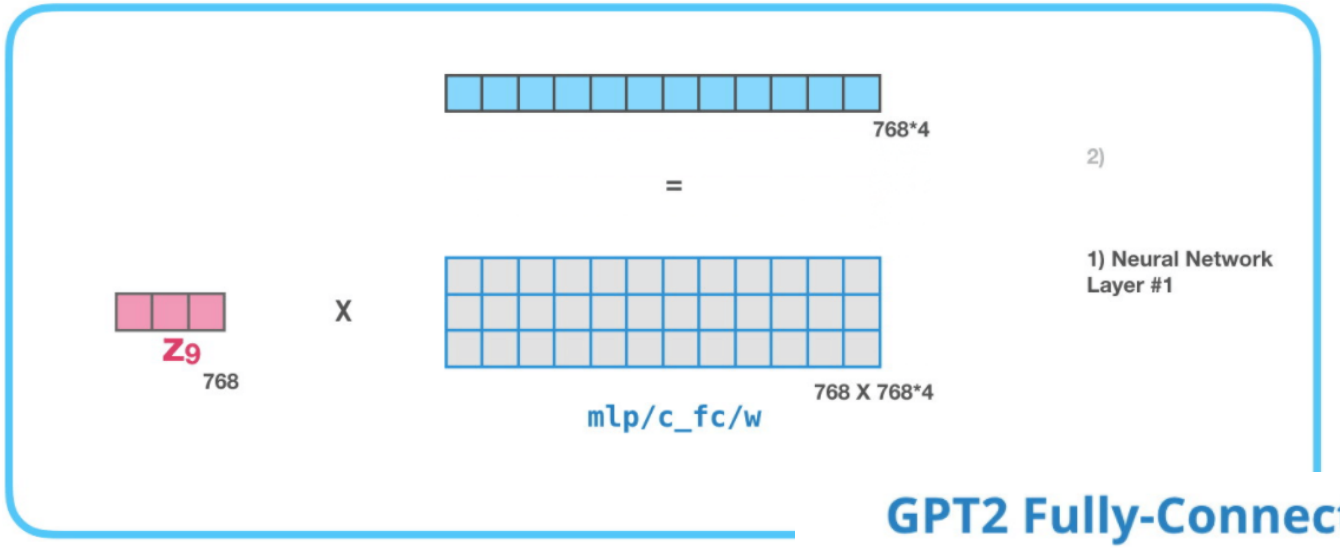




1) Create Q,K,V
1.5) Split attention head
2) Score (Q랑 K곱해서)
3) SUM (Z구함)
3.5) Merge attention heads
4) 구해진 z에 projection

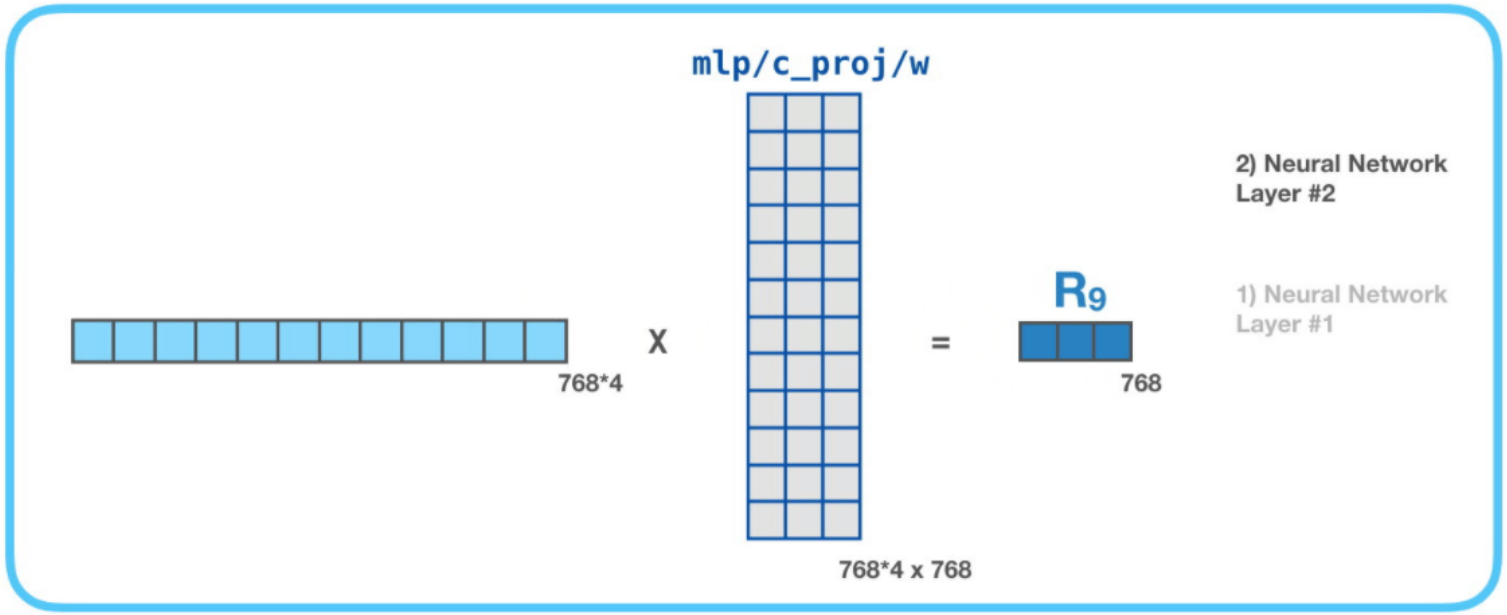


GPT2 Fully-Connected Neural Network

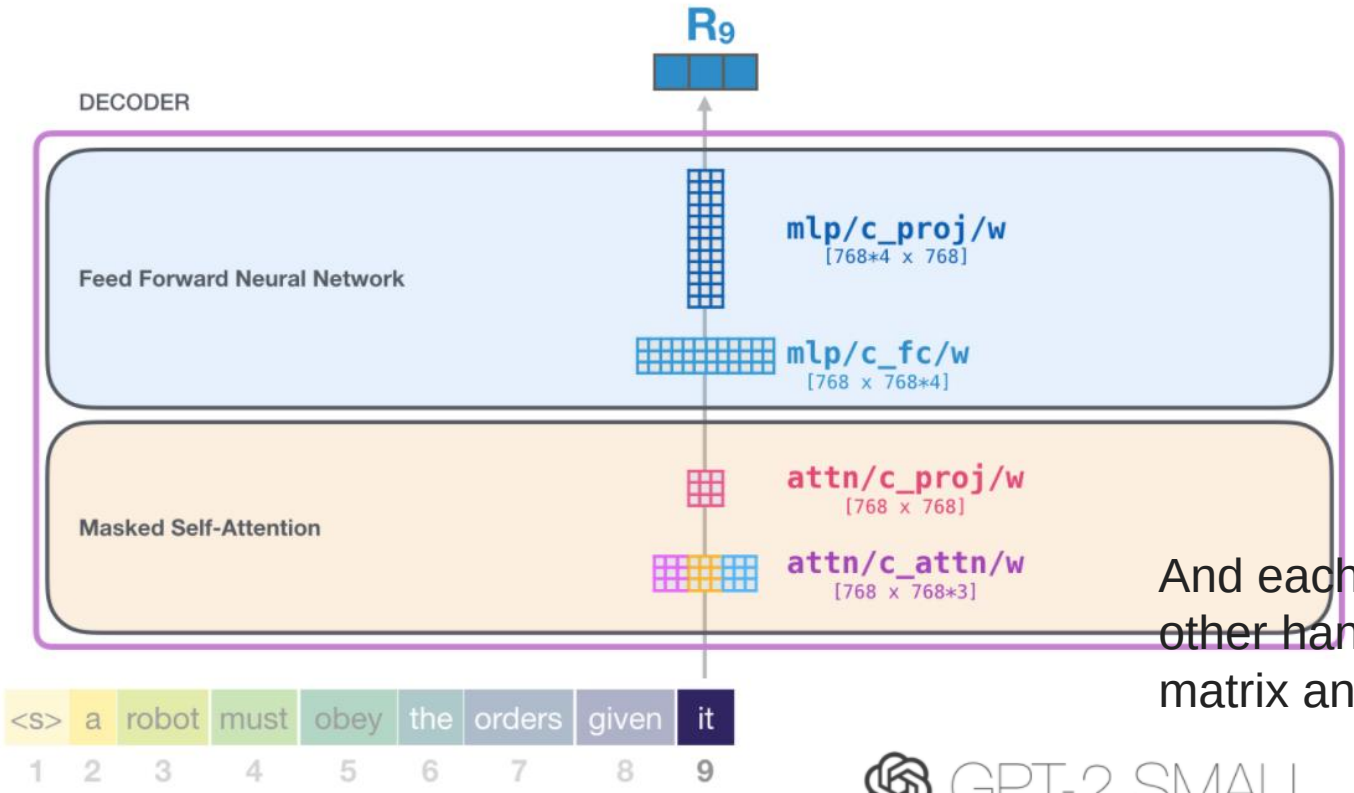


Feed-Forward network에서
1) 받은 Z에 weight곱해서 펴주고
2) Projection해서 인풋이랑 dim맞춘 R출력

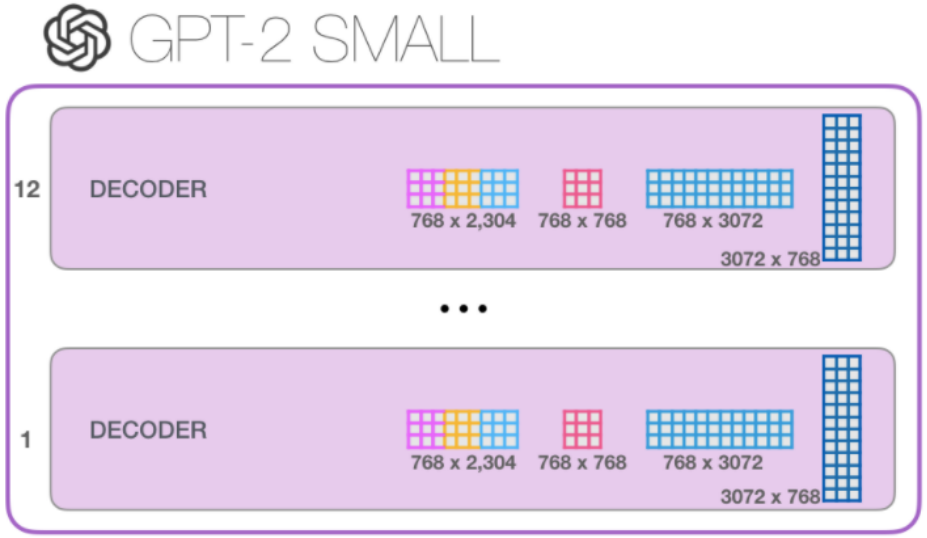
GPT2 Fully-Connected Neural Network

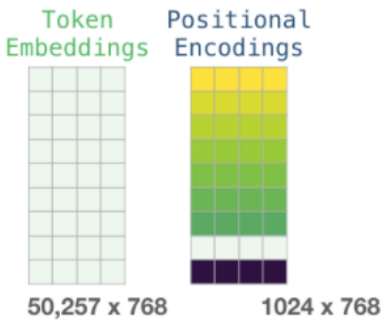


(Not shown: A bias vector)



And each block has its own set of these weights. On the other hand, the model has only one token embedding matrix and one positional encoding matrix:





Architecture

- modified initialization, pre-normalization, reversable tokenization 적용
- 단, 트랜스포머 레이어의 attention 패턴에 대해 dense와 locally banded sparse attention을 번갈아 사용
- 스케일에 따라 다음과 같이 8가지 모델을 학습하고 테스트. 가장 큰 모델은 96층의 레이어, 12,288차원의 히든 차원, 96개 attention head를 가지는 총 1750억 개의 파라미터의 모델임. 모든 모델은 3,000억 토큰에 대해 학습함.

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}

- 더 큰 모델에 대해서는 더 큰 배치를 적용했으나, 오히려 learning rate는 작게 적용함.
- 학습 과정에서 그라디언트의 noise scale을 측정해 배치 사이즈를 정하는 데에 활용. (관련 연구)
- 큰 모델 학습에는 메모리가 부족하기 때문에, 행렬 곱에 있어 모델 병렬화와 레이어 사이의 모델 병렬화를 섞어서 사용
- 더 자세한 훈련 과정은 Appendix B 참고...

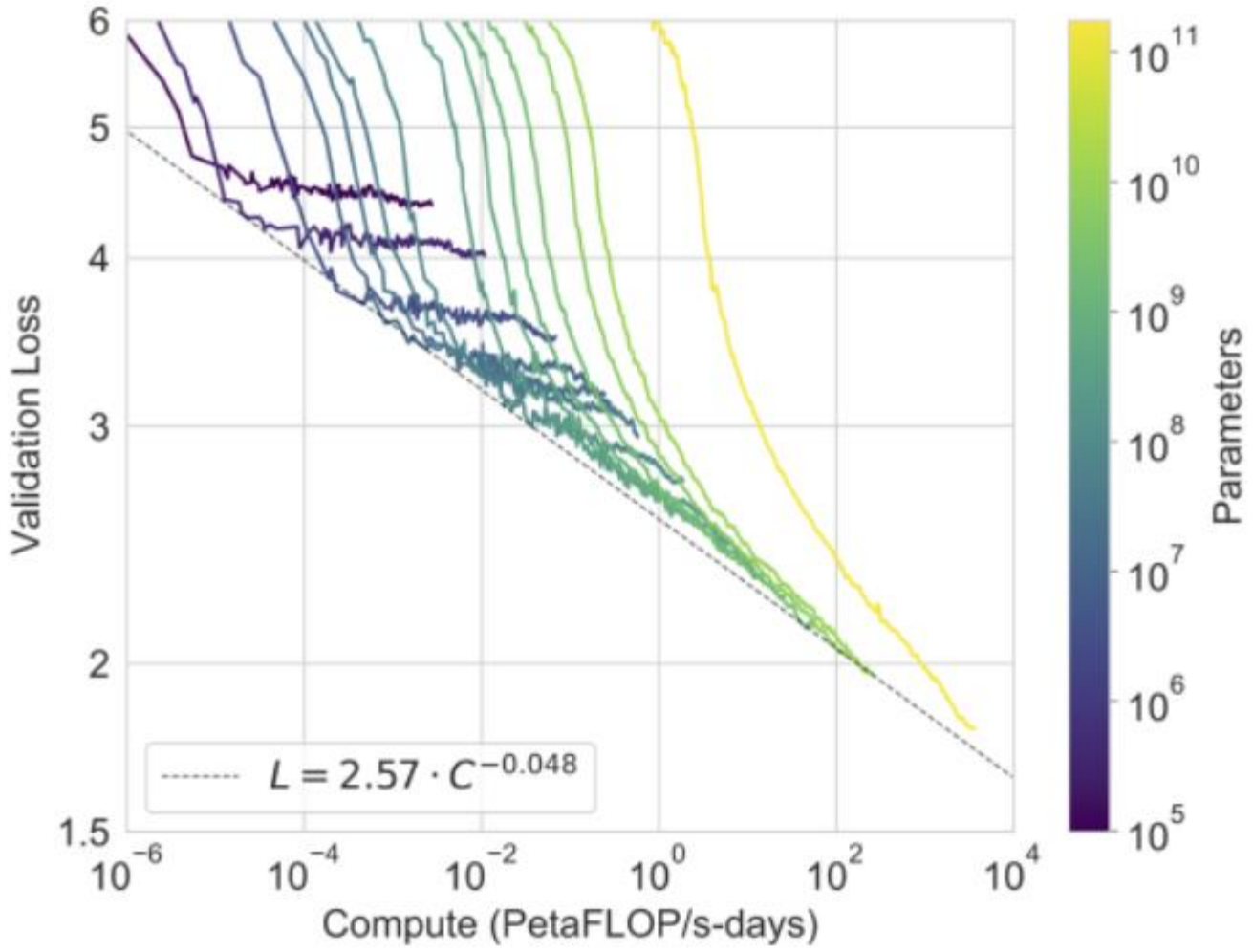
Datasets

- Training is the process of exposing the model to lots of text.(300billion; 300,000,000,000, 3000억; tokens of text) That process has been completed. All the experiments you see now are from that **one trained model.**

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.

Results



다음 그림은 스케일에 따른 8개의 모델에 대한 훈련 곡선을 나타낸다. 이 그래프에서는 십만 개의 파라미터만을 가진 극단적으로 작은 6개의 모델 결과도 보여준다.

파라미터 많을 수록 Loss 떨어짐
계산 속도는 (x축) 늘어나고

3.1 전통적인 language modeling 및 관련 태스크에서의 성능

Setting	LAMBADA (acc)	LAMBADA (ppl)	StoryCloze (acc)	HellaSwag (acc)
SOTA	68.0[a]	8.63[b]	**91.8**[c]	**85.6**[d]
GPT-3 Zero-Shot	**76.2**	**3.00**	83.2	78.9
GPT-3 One-Shot	**72.5**	**3.35**	84.7	78.1
GPT-3 Few-Shot	**86.4**	**1.92**	87.7	79.3

Language Modeling
: Penn Tree Bank 데이터에 대해 zero-shot perplexity를 계산한 결과, 가장 큰 GPT-3 모델은 기존 zero-shot SOTA보다 15 point 앞선 20.50의 perplexity로 SOTA를 달성함
LAMBADA (문장 완성하기/ 언어의 장기 의존성을 모델링하는 태스크)

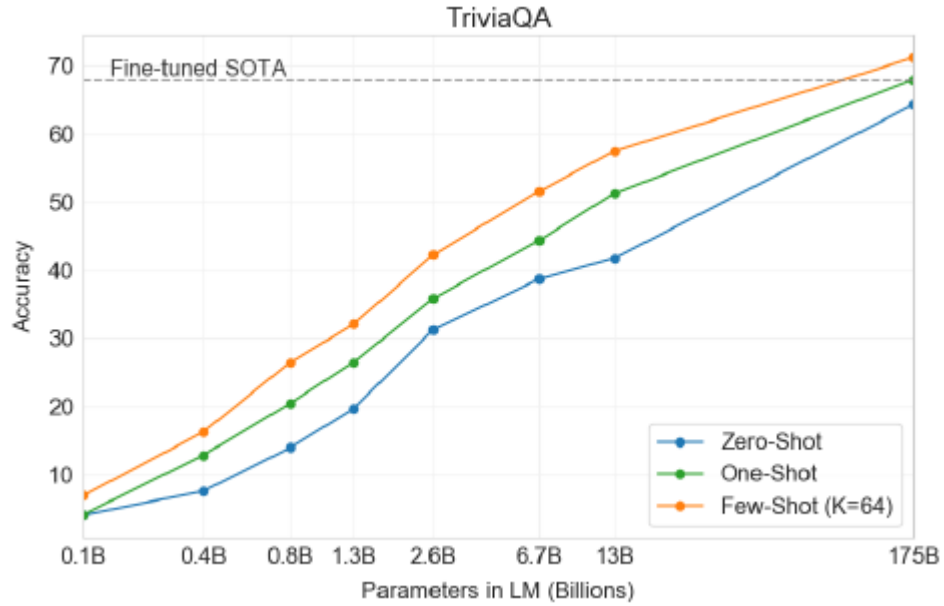
Alice was friends with Bob. Alice went to visit her friend _____. → Bob

George bought some baseball equipment, a ball, a glove, and a _____. →

HellaSwag (짧은 글이나 지시사항을 끝맺기에 가장 알맞은 문장을 고르는 태스크)
: 모델은 어려워하지만 사람에게는 쉬운 태스크 중 하나, 현 SOTA인 multi-task 학습 후 fine-tuning 전략을 취한 ALUM 에는 미치지 못하는 성능을 얻었다.

3.2 Closed Book Question Answering

Setting	NaturalQS	WebQS	TriviaQA
RAG (Fine-tuned, Open-Domain) [LPP+20]	**44.5**	**45.5**	**68.0**
T5-11B+SSM (Fine-tuned, Closed-Book) [RRS20]	36.6	44.7	60.5
T5-11B (Fine-tuned, Closed-Book)	34.5	37.4	50.1
GPT-3 Zero-Shot	14.6	14.4	64.3
GPT-3 One-Shot	23.0	25.3	**68.0**
GPT-3 Few-Shot	29.9	41.5	**71.2**

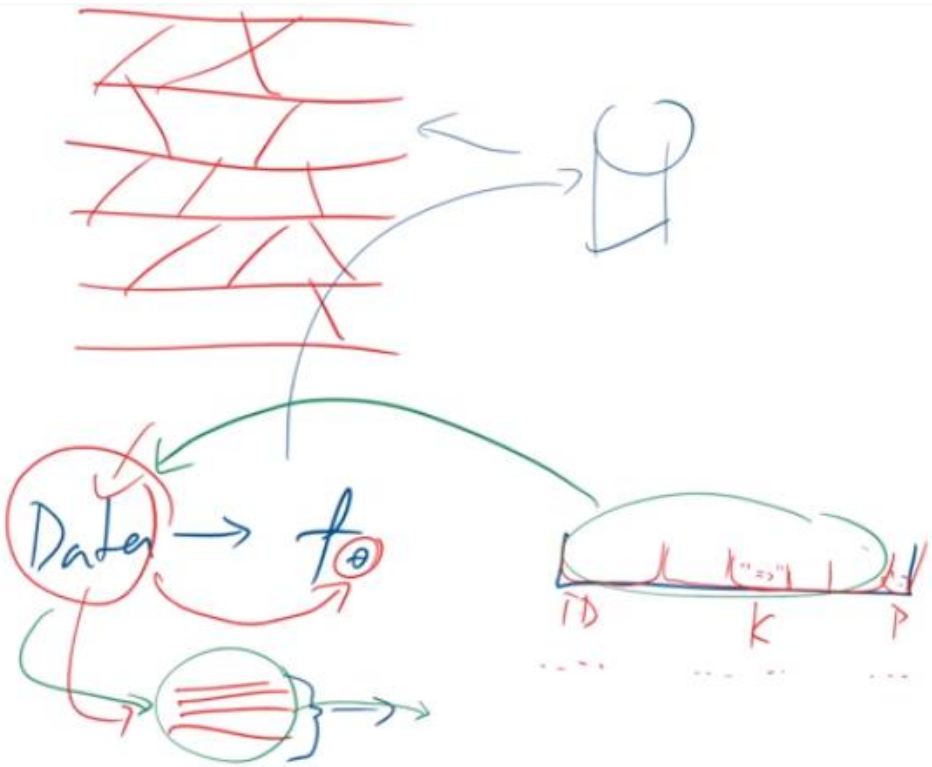


TriviaQA : Few-shot& Zero-shot 성능으로 T5-11B 모델의 **fine-tuning 기반의 접근법 성능을 뛰어넘음**

WebQuestions : 14.4% / 25.3% / 41.5% (0S/1S/FS) -> few shot으로 갔을 때 zero shot에 비해 성능 향상이 큰 태스크 중 하나. GPT-3에게 있어 **WebQA스타일의 질문은 out-of-distribution이었을 것으로 추정함**. 그럼에도 불구하고 T5-11B fine-tuning 전략 성능인 37.4%을 넘었고, Q&A를 위한 사전학습을 더한 T5-11B+SSM의 44.7% 성능에 비견할 만하다.

Natural Questions : 14.6% / 23.0% / 29.9% -> NQ는 위키피디아에 대해 **fine-grained 지식**을 요구하는 태스크로, GPT-3의 폭넓은 사전학습 분포에 대한 학습 능력을 측정하기에 적당하지 않았다고 분석함.

OPINION



Simply storing the training data in the connection of Giant Network

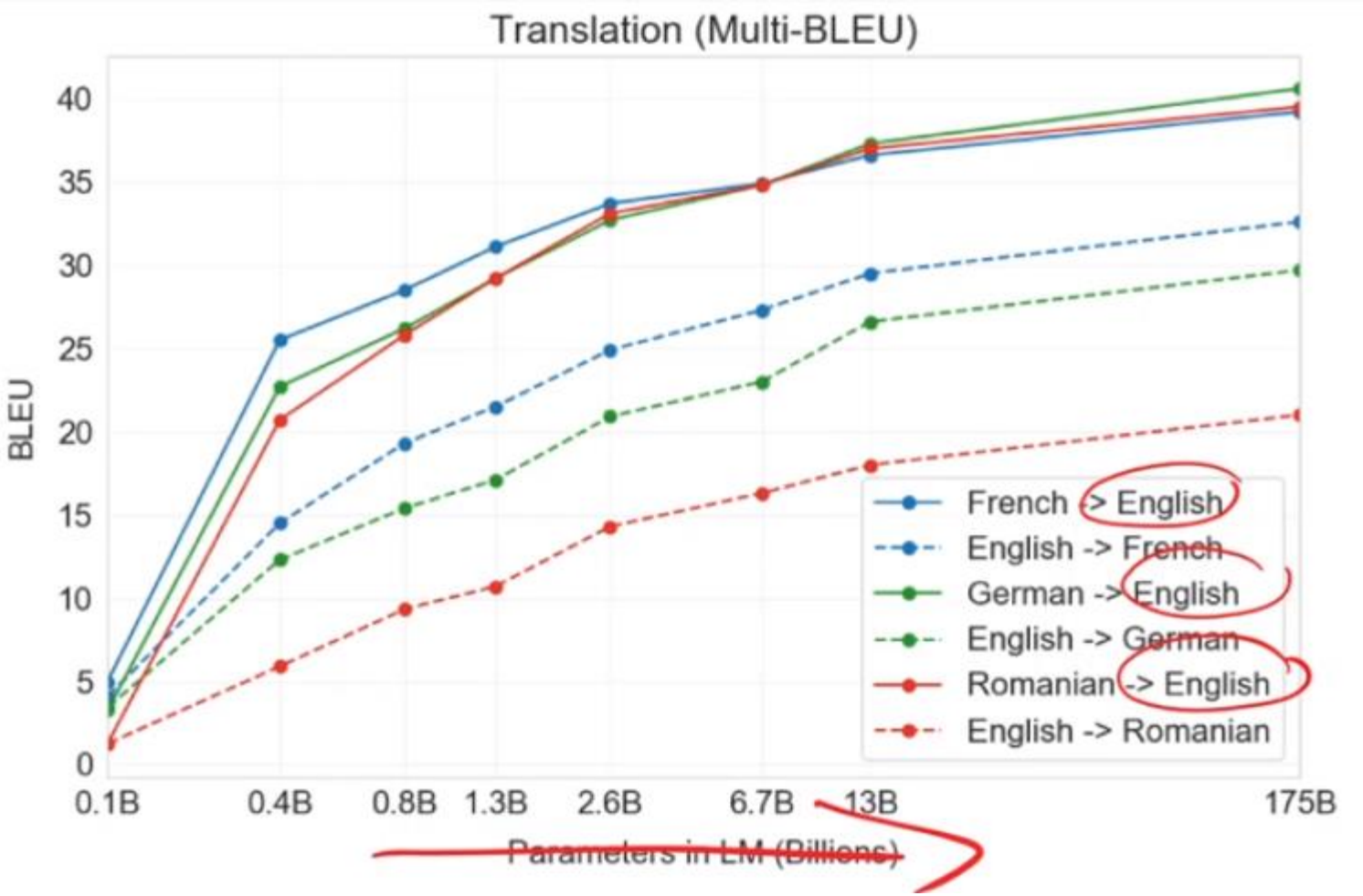
Data를 가지고 DL의 parameter를 학습시킴 (supervised learning) 결국 Data가 Parameter에 영향을 미친다.

GPT-3에 1700억개의 parameter를 가지고 있으니, 그 parameter마다 효율적으로 training data를 storing (저장)하고 query받으면 **Interpolation**해서 답을 내는 것이 아니냐 .

질의 자체가, ID + K 개의 샘플 + 마지막에 prompt (질의) 로 이루어져 있으므로 ID+K샘플을 Training data에서 찾아서 (weight)를 그리고 나서 pattern matching (interpolation)을 통해서 답이 나오지 않나?

Reasoning이 있다고 생각하진 않는다.

3.3 Translation



Target Languate가 English일 때가 높다. 그러나 다른 언어가 Target이면 낮다.
Training 데이터가 English 가 더 많을 테니까

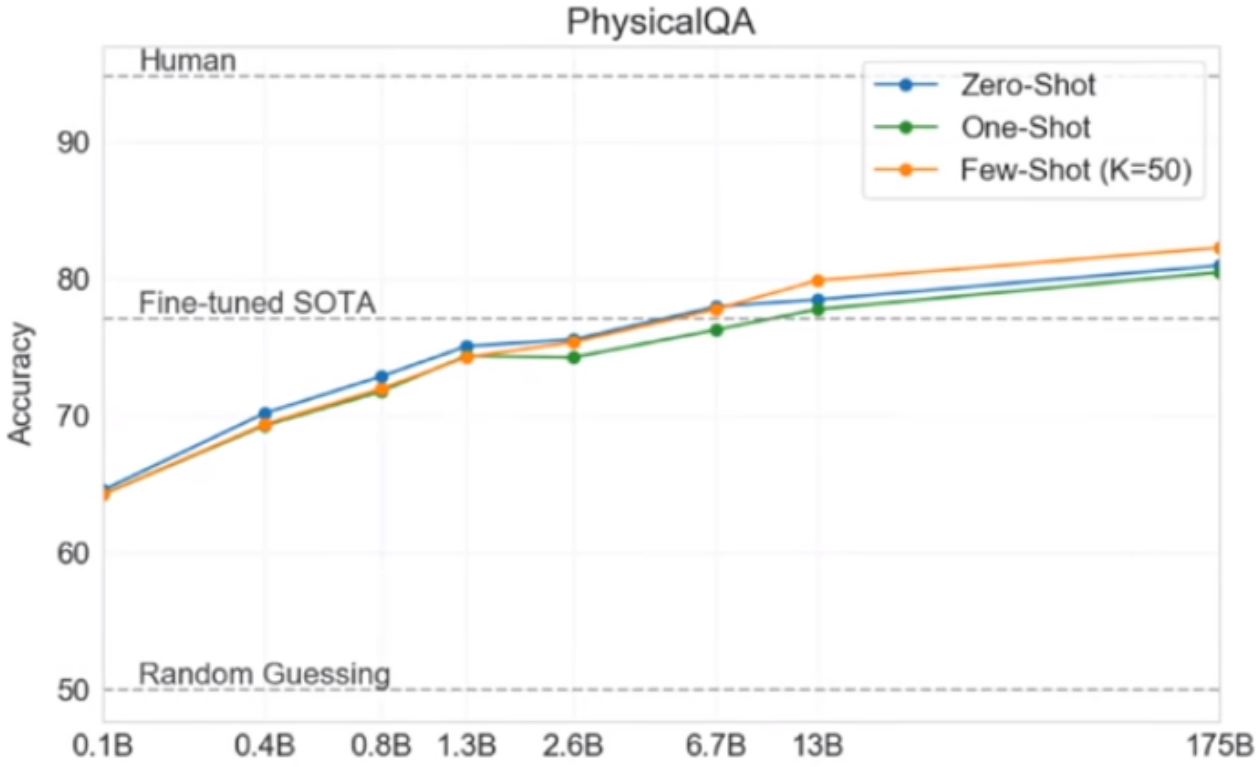
Web에서 두개 언어가 있는 데를 찾고 (France, English) 그 데이터를 가지고 그럴싸하게 번역한다

3.4 대명사 지칭 문제 (Winograd-syle태스크)

Setting	Winograd	Winogrande (XL)
Fine-tuned SOTA	**90.1**[a]	**84.6**[b]
GPT-3 Zero-Shot	88.3*	70.2
GPT-3 One-Shot	89.7*	73.2
GPT-3 Few-Shot	88.6*	77.7

Fine-tuned RoBERTa-Large랑 비슷함.

3.5 Common Sense Reasoning



물리학이 어떻게 작동하는지 묻는 PhysicalQA(PIQA)에서는 few / zero shot 세팅에서 이미 SOTA를 넘겼지만, 분석 결과 **data contamination issue**가 있을 수 있다고 조사함. PIQA데이터 셋이랑 Train데이터 가 겹침을 확인 (너무 오래 트레이닝 해서 못뺌)

3-9학년 과학 시험 수준의 4지선다형 문제인 ARC에서는 easy와 challenge 모두 SOTA에 미치지 못하는 성적. OpenBookQA에서는 few-shot이 zero-shot setting 대비 크게 성능 향상이 있어 in-context learning을 해낸 것으로 보이나, 역시 SOTA에는 미치지 못하는 성적이었다.

3.6 기계 독해

Setting	CoQA	DROP	QuAC	SQuADv2	RACE-h	RACE-m
Fine-tuned SOTA	**90.7**[a]	**89.1**[b]	**74.4**[c]	**93.0**[d]	**90.0**[e]	**93.1**[e]
GPT-3 Zero-Shot	81.5	23.6	41.5	59.5	45.5	58.4
GPT-3 One-Shot	84.0	34.3	43.3	65.4	45.9	57.4
GPT-3 Few-Shot	85.0	36.5	44.3	69.8	46.8	58.1

- SOTA보다 낮음
- 5 datasets abstraction, multiple choice, span based answer formats in both dialog and single question settings.
- 문단을 읽고, 문제에 답하는 문제 -> 긴 문단을 Training data에서 찾아서 pattern matching을 하기 어렵다. 실제 reasoning을 의미하므로 그래서 낮은게 아닐까?

3.7 SuperGLUE

- SOTA보다는 낮음, BERT보단 조금 높음
- .

	SuperGLUE Average	BoolQ Accuracy	CB Accuracy	CB F1	COPA Accuracy	RTE Accuracy
Fine-tuned SOTA	**89.0**	**91.0**	**96.9**	**93.9**	**94.8**	**92.5**
Fine-tuned BERT-Large	69.0	77.4	83.6	75.7	70.6	71.7
GPT-3 Few-Shot	71.8	76.4	75.6	52.0	92.0	69.0

	WiC Accuracy	WSC Accuracy	MultiRC Accuracy	MultiRC F1a	ReCoRD Accuracy	ReCoRD F1
Fine-tuned SOTA	**76.1**	**93.8**	**62.3**	**88.2**	**92.5**	**93.3**
Fine-tuned BERT-Large	69.6	64.6	24.1	70.0	71.3	72.0
GPT-3 Few-Shot	49.4	80.1	30.5	75.4	90.2	91.1

The BoolQ dataset release consists of three `.jsonl` files, where each line is a JSON dictionary with the following format:

```
{
  "question": "is france the same timezone as the uk",
  "passage": "At the Liberation of France in the summer of 1944, Metropolitan France kept GMT+2 as it was the time then
  "answer": false,
  "title": "Time in France",
}
```

BoolQ는 Reasoning에 가까움
그래서 못함

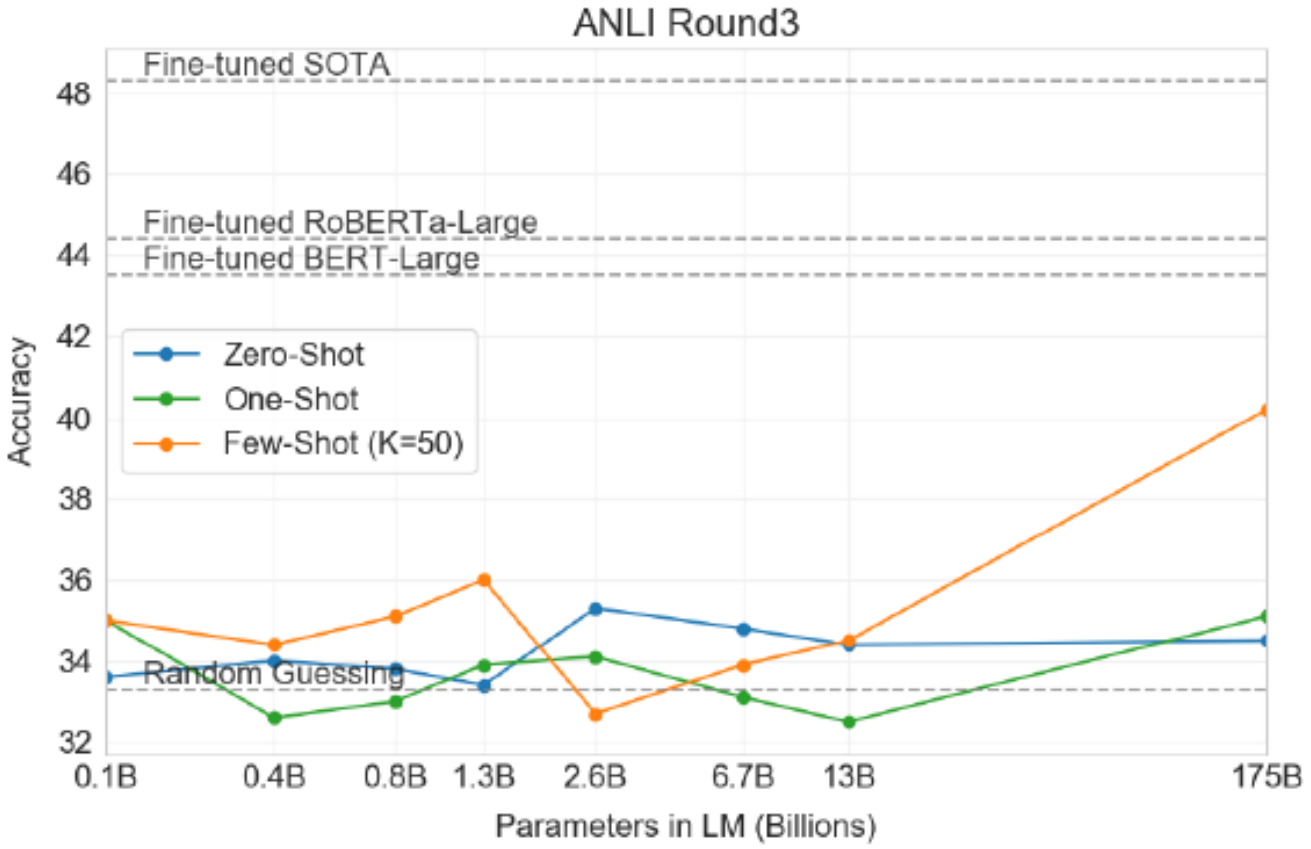
1. Examples

Premise: The man broke his toe. What was the CAUSE of this?
Alternative 1: He got a hole in his sock.
Alternative 2: He dropped a hammer on his foot.

COPA 는 질문에 해당하는 답을 선택, 이건 Training data 중에서 broke toe와 hammer가 더 많이 등장했을 것이다.
LM이 잘 할 수 있는 일이다.

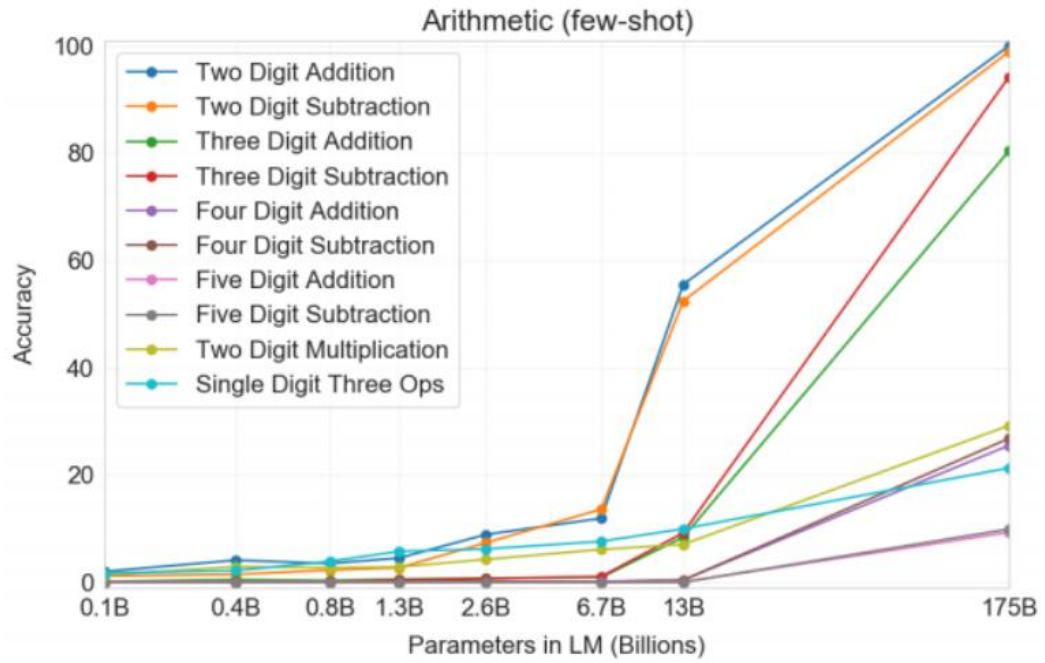
3.8 NLI

Natural Language Inference (NLI) [Fyo00] concerns the ability to understand the relationship between two sentences. In practice, this task is usually structured as a two or three class classification problem where the model classifies whether the second sentence logically follows from the first, contradicts the first sentence, or is possibly true (neutral).



이 task도 reasoning의 일환으로 SOTA에 한참 못 미침

3.9 Synthetic & Qualitive Tasks



2자리, 3자리, 4자리, 5자리 더하기, 빼기, 곱하기 등

"Q: What is 48 plus 76? A: 124."

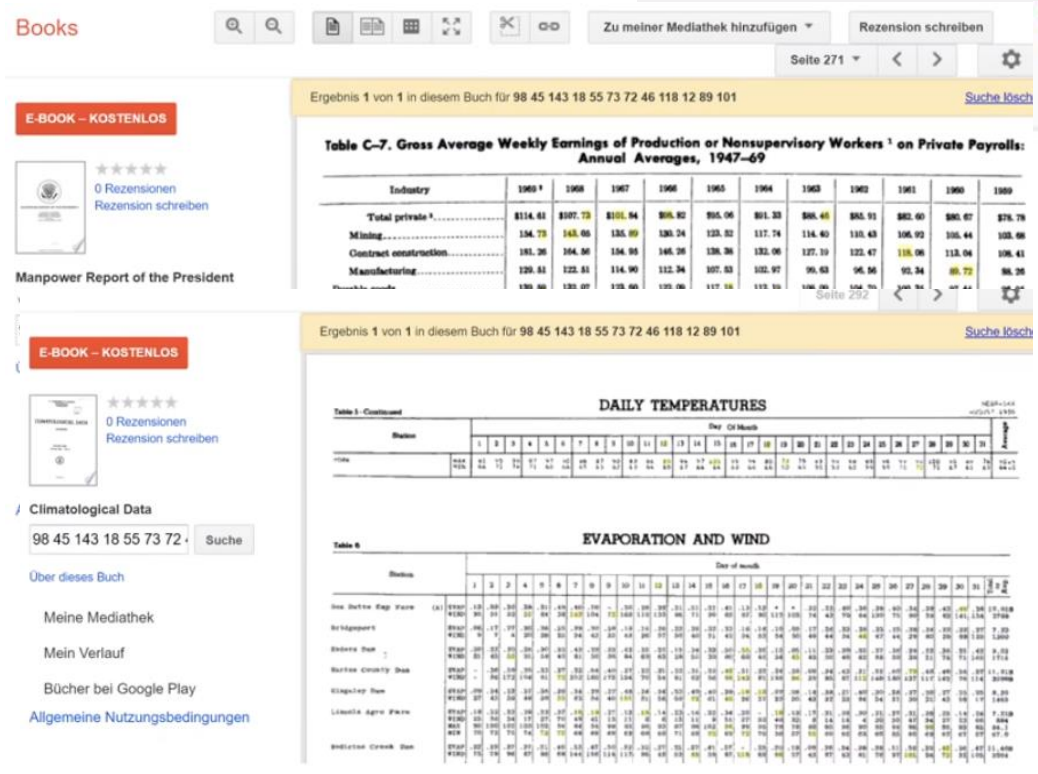
문장을 계산하니까 reasoning을 한다고 생각할 수 있지만,
결과를 보면 param이 작으면 결과가 안좋고, param이 커지면 2자리 덧셈, 뺄셈은 90% 정확도를 나타낸다.
하지만 3,4자리는 낮다 (논리적으로 어려워서??그럴리가 덧셈이란 논리는 같은데)

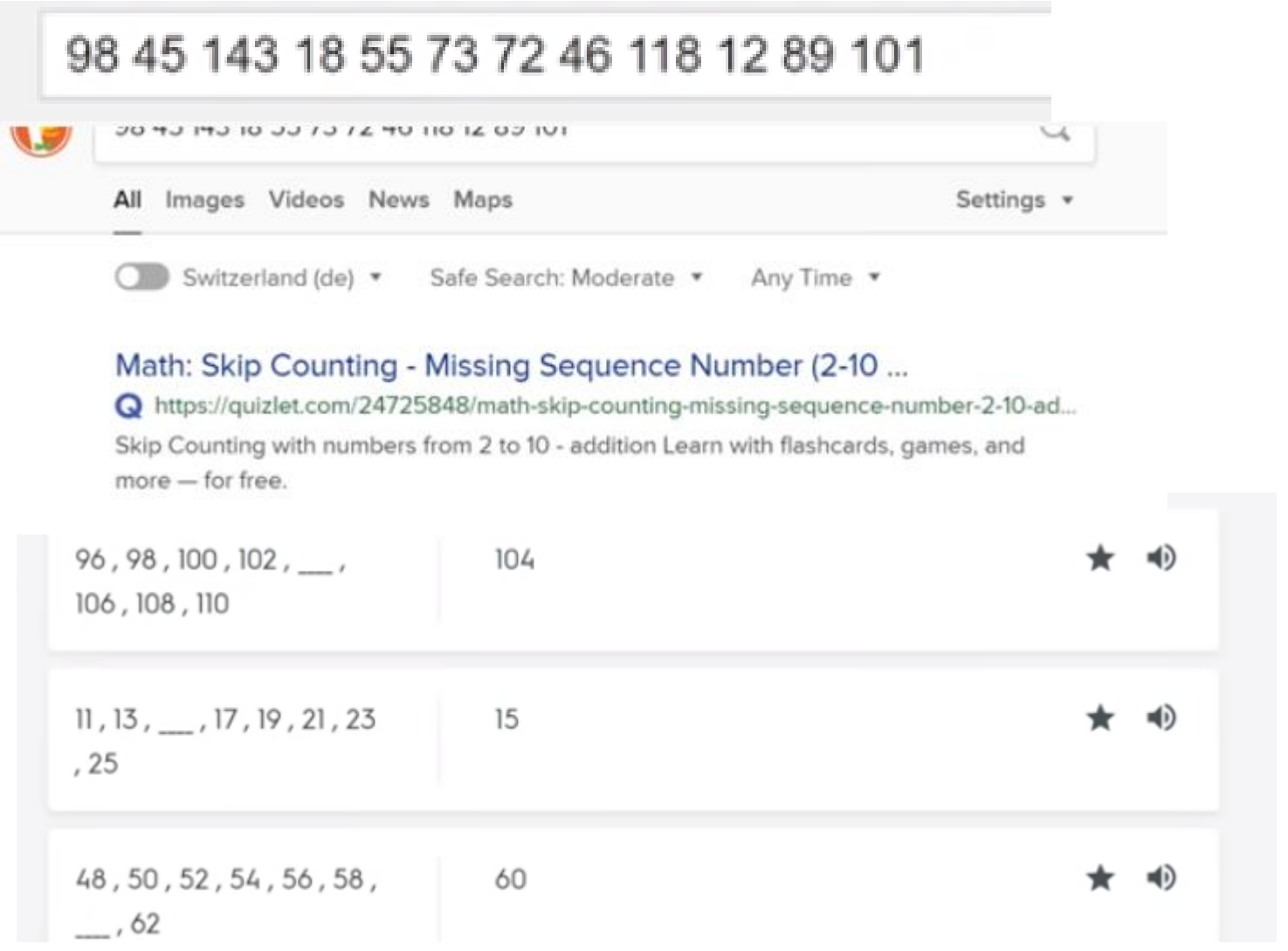
두자리 수가 네자리, 다섯자리수보다 데이터가 많으니까

몇 개의 예제를 알려주면, 인터넷에서 그 숫자만 찾아도 결과를 알아 낼 수 있다.

3.9 Synthetic & Qualitive Tasks

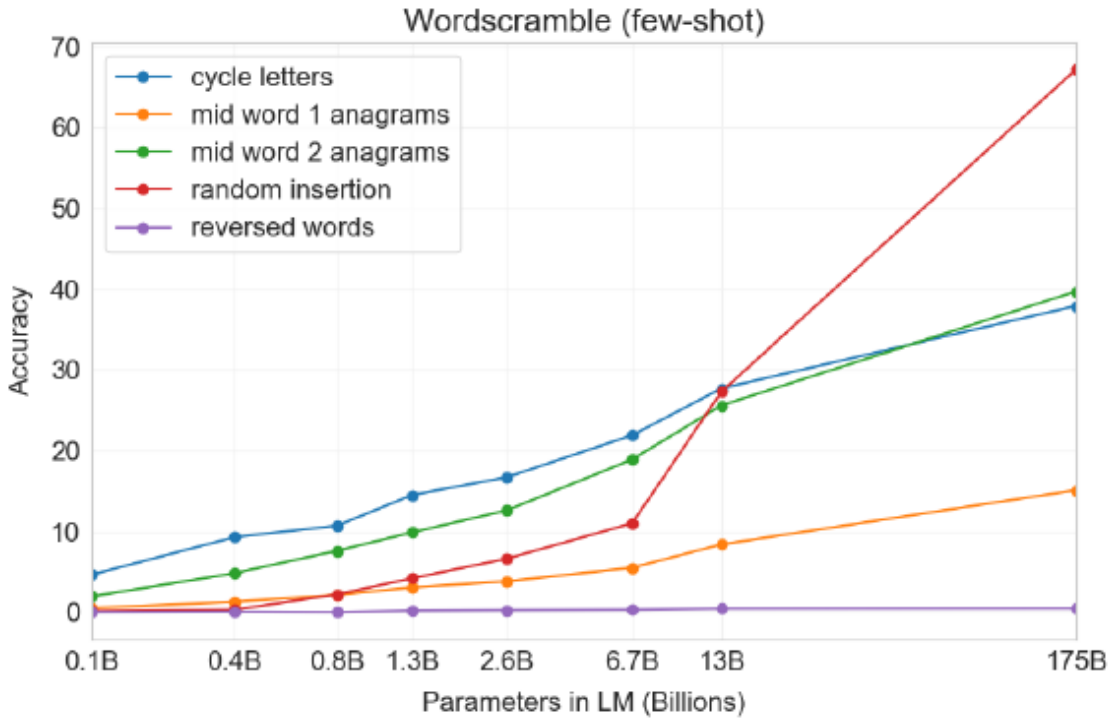
몇 개의 예제를 알려주면, 인터넷에서 그 숫자만 찾아도 결과를 알아 낼 수 있다.





https://youtu.be/SY5PvZrJhLE?t=2749

3.9 Word Scrambling & Manipulation



- CL(Cycle letters in word): 순서가 뒤죽박죽 된 단어를 원래로 맞추기 (예. plepa -> apple)
- A1(Anagram, 처음/마지막 글자 제외) : 처음과 마지막 글자를 제외하고 랜덤하게 만든 글자 원복하기 (aekve -> apple)
- A2(Anagram, 처음&마지막 두 글자 제외) : 처음&마지막 두 글자씩을 제외하고 랜덤하게 만든 글자 원복 (apyle -> apple)
-RI(Random insertation in word) : 각 글자 사이에 랜덤한 마침표나 띄어쓰기가 들어간 것 원복 (a!p.p l^e% -> apple)
- RW(Reversed word) : 거꾸로 쓴 글자를 원래로 맞추기 (elppa -> apple)

제대로 된 단어로 바꾸는 RI (Random Insertion), CL (Cycle letters)는 잘함 “lyinevitab” -> “inevitably” //
그리고 모든 task가 영단어로 되어 있음. 영어 단어로 training한 점을 생각할 때 제대로 된 단어를 찾아내는 것은 쉽다. 하려면 bilingual로 scrambling해봤어야
정답 단어 자체가 training data랑 겹쳤다. (Word Pieces) simply calling training data의 역할일 뿐
또 해보려면 정상 단어에서 이상한 단어로 만드는 것을 해봤어야 한다.

3.10 SAT

audacious is to boldness as
(a) sanctimonious is to hypocrisy,
(b) anonymous is to identity,
(c) remorseful is to misdeed,
(d) deleterious is to result,
(e) impressionable is to temptation

대담함이란 대담함을 의미합니다.
(A) 신성함은 위선이며,
(B) 익명은 정체성에,
(C) 후회는 잘못하는것,
(D) 해로운 것은 결과에,
(E) 감명은 유혹

뭔소리야, 여튼 동의어 찾는 TASK

GPT-3은 53.7/59.1/65.2%(K=20)의 정확도를 보임. 대학생 평균이 57%인 것에 비하면 GPT-3은 단어 사이의 관계를 잘 학습했다!

근데 원래 LM 모델이 동의어 찾기에 능통 , word relation 에 능함 (단어간 확률 학습하니까)

3.11 News Generation

각기 다른 사이즈의 GPT-3 모델이 생성한 200 단어 미만의 짧은 뉴스가 "사람이 생성한 것인지, 기계가 생성한 것인지" 사람이 평가해보는 세팅에서 가장 큰 모델은 175B모델의 경우 **평균 52%의 정확도를** 보였다. 기계가 생성한 글을 기계가 생성했다고 **판별하기 어려운 수준.**

	Mean accuracy	95% Confidence Interval (low, hi)	t compared to control (p-value)	"I don't know" assignments
Control (deliberately bad model)	86%	83%–90%	-	3.6 %
GPT-3 Small	76%	72%–80%	3.9 (2e-4)	4.9%
GPT-3 Medium	61%	58%–65%	10.3 (7e-21)	6.0%
GPT-3 Large	68%	64%–72%	7.3 (3e-11)	8.7%
GPT-3 XL	62%	59%–65%	10.7 (1e-19)	7.5%
GPT-3 2.7B	62%	58%–65%	10.4 (5e-19)	7.1%
GPT-3 6.7B	60%	56%–63%	11.2 (3e-21)	6.2%
GPT-3 13B	55%	52%–58%	15.3 (1e-32)	7.1%
GPT-3 175B	52%	49%–54%	16.9 (1e-34)	7.8%

3.11 News Generation

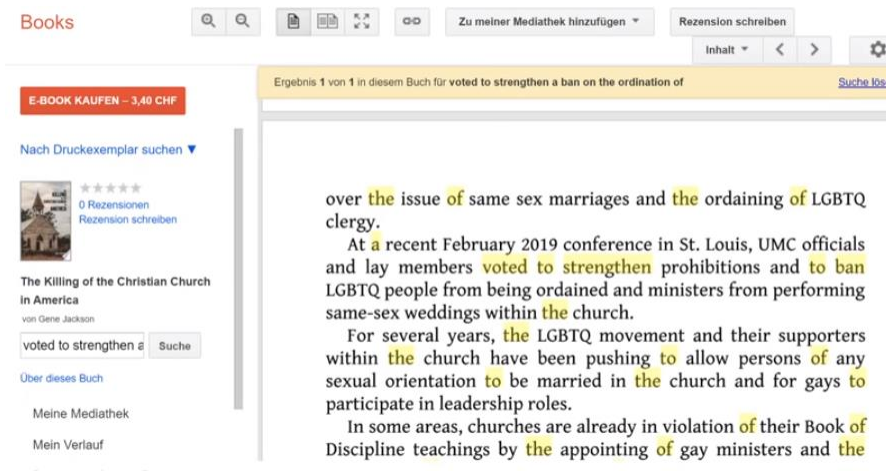
Title: United Methodists Agree to Historic Split
Subtitle: Those who oppose gay marriage will form their own denomination
Article: After two days of intense debate, the United Methodist Church has agreed to a historic split - one that is expected to end in the creation of a new denomination, one that will be "theologically and socially conservative," according to The Washington Post. The majority of delegates attending the church's annual General Conference in May voted to strengthen a ban on the ordination of LGBTQ clergy and to write new rules that will "discipline" clergy who officiate at same-sex weddings. But those who opposed these measures have a new plan: They say they will form a separate denomination by 2020, calling their church the Christian Methodist denomination.
The Post notes that the denomination, which claims 12.5 million members, was in the early 20th century the "largest Protestant denomination in the U.S.," but that it has been shrinking in recent decades. The new split will be the second in the church's history. The first occurred in 1968, when roughly 10 percent of the denomination left to form the Evangelical United Brethren Church. The Post notes that the proposed split "comes at a critical time for the church, which has been losing members for years," which has been "pushed toward the brink of a schism over the role of LGBTQ people in the church." Gay marriage is not the only issue that has divided the church. In 2016, the denomination was split over ordination of transgender clergy, with

1. 모델이 Language를 너무 잘 익혀서 context에 맞춰서 writing을 잘하는거다.

2. 주어진 title, subtitle가지고 training data에서 pattern matching해서 interpolation해서 뉴스 만드는게 아니냐?

저자들은 news article이 training 데이터에 없는것을 확인해서 1번이 맞다고 주장했다.

근데 아무 문장이나 검색해서 찾으면 "vote to strengthen …" 같은 문제 LGBT 에 대해서 말한 비슷한 문장들을 쉽게 검색할 수 있다. -> 그러므로 2번 주장



3.12 문법 교정

A "Burringo" is a car with very fast acceleration. An example of a sentence that uses the word Burringo is:
In our garage we have a Burringo that my father drives to work every day.

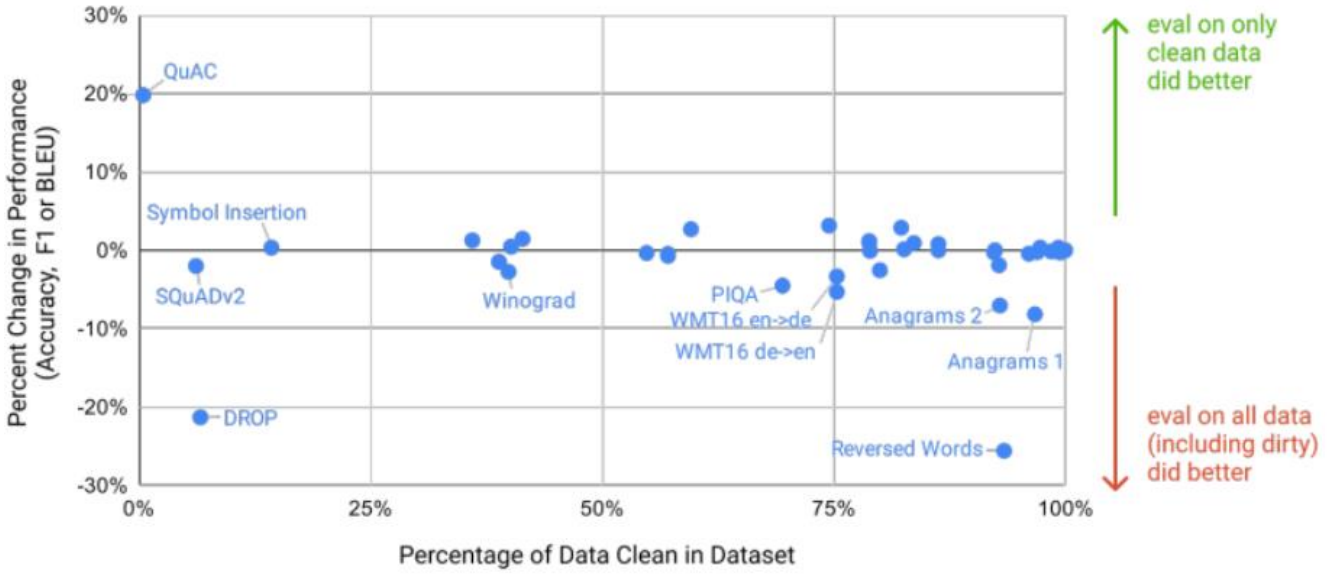
A "Gigamuru" is a type of Japanese musical instrument. An example of a sentence that uses the word Gigamuru is:
I have a Gigamuru that my uncle gave me as a gift. I love to play it at home.

To "screeg" something is to swing a sword at it. An example of a sentence that uses the word screeg is:
We screeghed at each other for several minutes and then we went outside and ate ice cream.

문제에 주어진 영어 단어가 다 주어져있고, 영어 기본이고 다른 언어는 잘 되지 않음.
Good -> Poor로 할 수는 없음 (training에 적으니까). 언어를 understanding했다면 Good에서 Poor도 만들 수 있어야 함.

LM으로써의 성능은 매우 좋음, 인공 일반 지능(Artificial general intelligence)까지는 아닌듯 Query 날리는 과정도 간단함.
검색엔진에 잘 활용될 수 있을 것 같음

GPT-3 사전학습 중 벤치마크 태스크를 외워버렸을 위험은 없는가



각각의 벤치마크에 대해 사전학습 데이터와 13-gram으로 **오버랩되는 데이터를 삭제하는 "클린"**버전의 테스트 셋을 만들어 보수적으로 모델을 평가하였다. 이후 GPT-3을 이러한 깨끗한 버전의 데이터에 대해 평가해보았을 때, 유출된 데이터에 대해 모델이 더 잘했다라는 특별한 증거는 없었다. 테스트셋 중 오염된 데이터의 비율이 높은 태스크에 대해서도 조사해 보았으나, 그것이 성능에 미치는 영향은 거의 0에 가까웠다.

1. 단순 n-gram 으로 duplication을 처리할 순 없다. New article에서 보았듯이 모든 단어가 다 일치하진 않지만 비슷한 topic을 나타내는 뉴스를 찾을 수 있었다. (fuzzy duplication & meaning duplication) 좋은 LM인건 맞지만 Reasoning 모델이라고 할 순 없다.

2. 각각의 데이터셋은 처리할 수 있지만, **데이터 셋들끼리도 Duplication** (In-between) 위키피디어 지식과 QA 지식과 겹침

Limitation

https://littlefoxdiary.tistory.com/44

1.성능적 한계

: 성능이 안좋은 애들 (reasoning부분들)

2. 모델의 구조/ 알고리즘적 한계

: 양방향적인 (bidirectional) 구조나 denoising 훈련 목적함수 등은 고려하지 않는다

3. 본질적인 한계 (fine-tune)

4. 훈련 과정의 효율성

: 사전학습을 위해서는 인간이 평생 보게 될 것보다 많은 양의 데이터

5. Few-shot 셋팅에서 효과의 불확실성

: few-shot learning이 정말로 추론 시에 새로운 태스크를 새롭게 배우는 것인지, 아니면 훈련하는 동안 배운 태스크 중 하나를 인지해 수행해내는 것인지는 모호하다.

6. 비용

7. 해석 가능성

: 데이터에 존재하는 편향(bias)