What is LLM and ChatGPT?

2023. 07. 28

Data Mining & Quality Analytics Lab.

채고은

발표자 소개

❖ 채고은 (Goeun Chae)

- 고려대학교 산업경영공학과 대학원 재학
- Data Mining & Quality Analytics Lab. (김성범 교수님)
- 석 · 박사 통합 과정 (2022. 03 ~ Present)

❖ Research Interest

- Self-Supervised Learning
- Hard Negative Sampling in Contrastive Learning
- Reinforcement Learning

❖ Contact

- goeunchae@korea.ac.kr

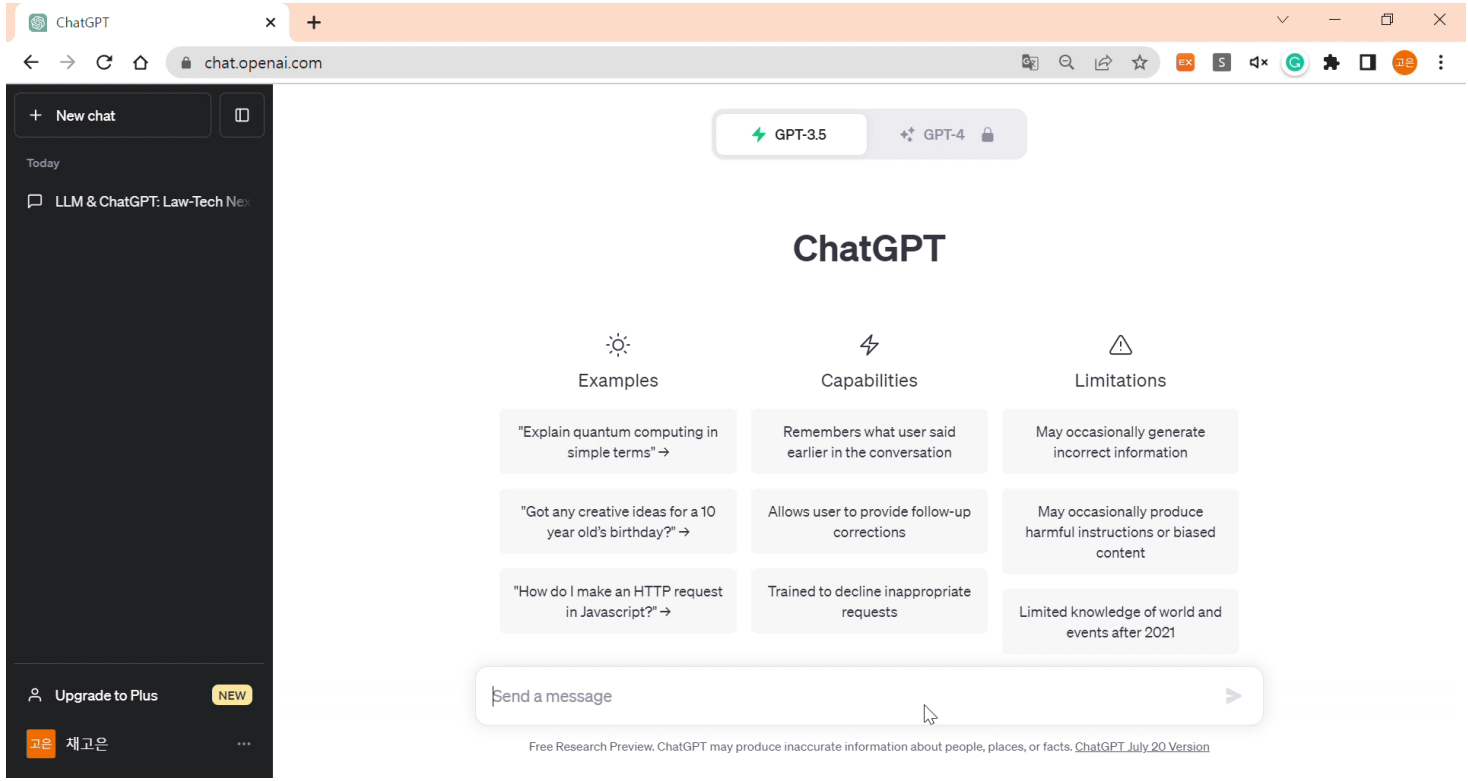
Contents

1. Introduction

Introduction

❖ Language Model

- 기계는 강력한 인공지능을 장착하지 않는 이상, 인간의 언어 형태로 이해하고 소통하는 것이 불가능
- 기계가 인간처럼 읽고, 쓰고, 소통할 수 있게 하는 기술



Introduction

❖ Language Modeling

- Language Modeling 의 목표: 단어 시퀀스의 생성 가능성을 모델링하여 미래 토큰의 확률 예측

History of Language Modeling

Neural Language Models	Pre-trained Language Models	Large Language Models
신경망을 사용해 단어 시퀀스의 확률을 특성화	신경망을 사전학습 한 후 Downstream 에 따라 미세조정	PLM에서 모델 & 데이터 크기를 확장
RNN	Transformer	ELMo 	GPT-2	LLaMa
LSTM	Seq2Seq	BERT 	GPT-3	Vicuna
GRU		GPT-1	Alpaca	Falcon

[1] Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223.

Introduction

❖ Language Modeling

- Language Modeling 의 목표: 단어 시퀀스의 생성 가능성을 모델링하여 미래 토큰의 확률 예측

History of Language Modeling

Neural Language Models

신경망을 사용해 단어 시퀀스의 확률을 특성화

- RNN
- LSTM
- GRU
- Word2Vec
- Seq2Seq

Pre-trained Language Models

신경망을 사전학습 한 후 Downstream 에 따라 미세조정

- Transformer
- ELMo 
- BERT 
- GPT-1

Large Language Models

PLM에서 모델 & 데이터 크기를 확장

- GPT-2
- GPT-3
- Alpaca
- LLaMa
- Vicuna
- Falcon

[1] Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223.

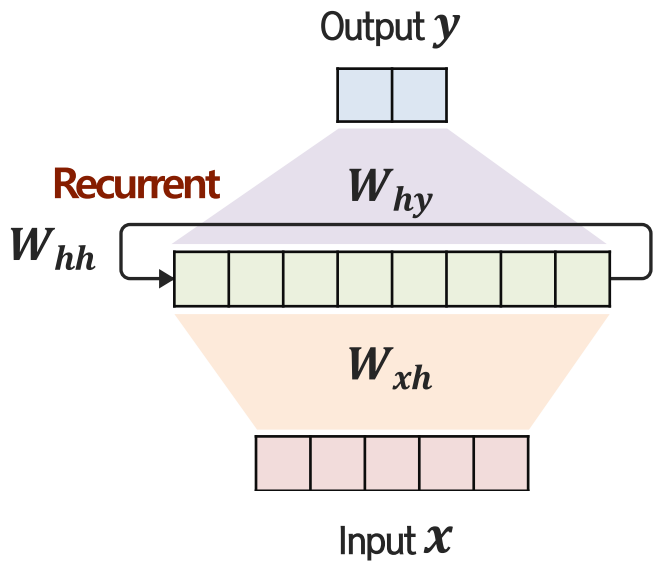
2. From Transformer to Large Language Model

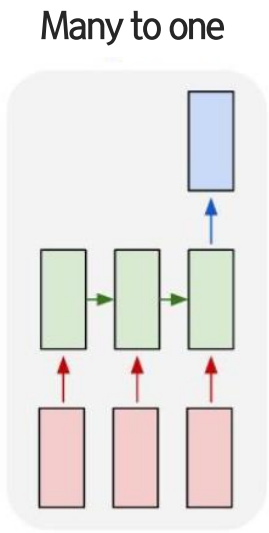
From Transformer to Large Language Model

Recurrent Neural Network (RNN)

❖ RNN Architecture

- RNN: 연속적인 데이터에서 연속된 변수들 간의 Dependency 를 반영
- 순차적으로 입력하고, 순차적으로 예측하는 알고리즘





Sentiment Classification

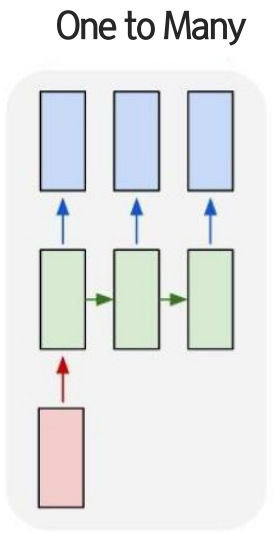
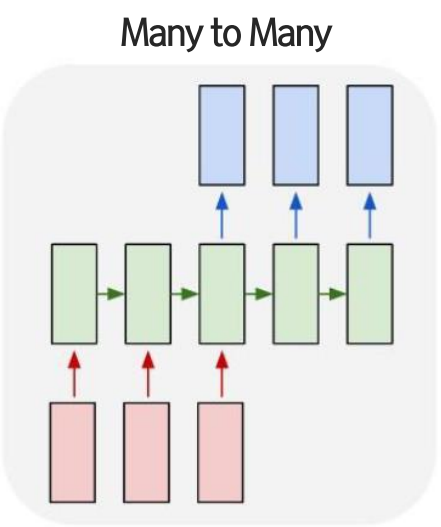


Image Captioning



Machine Translation

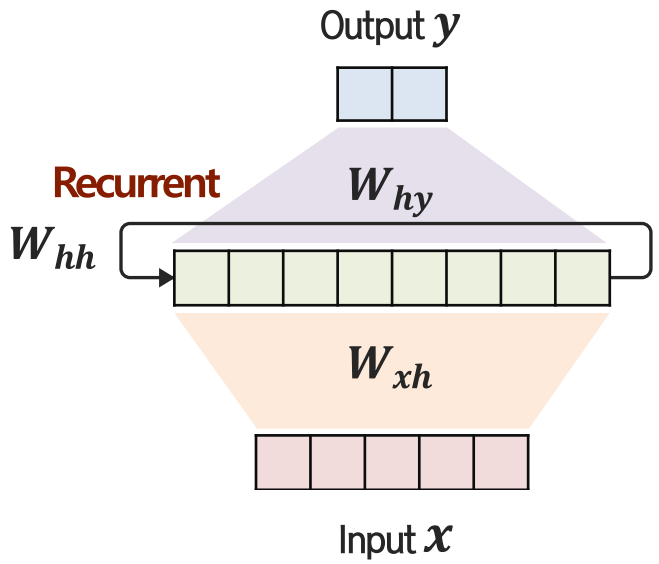
[2] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. Advances in neural information processing systems, 27.

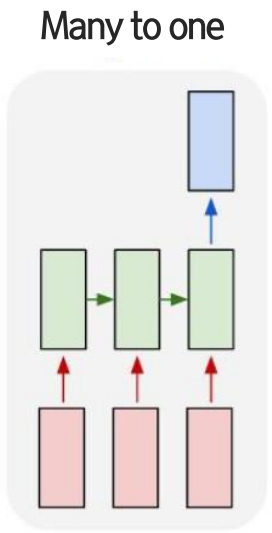
From Transformer to Large Language Model

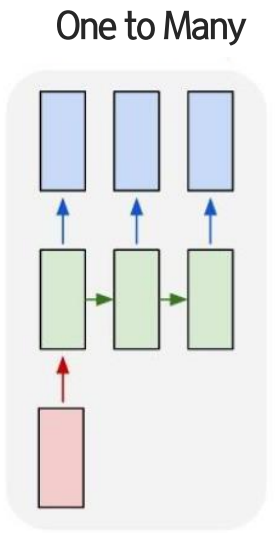
Recurrent Neural Network (RNN)

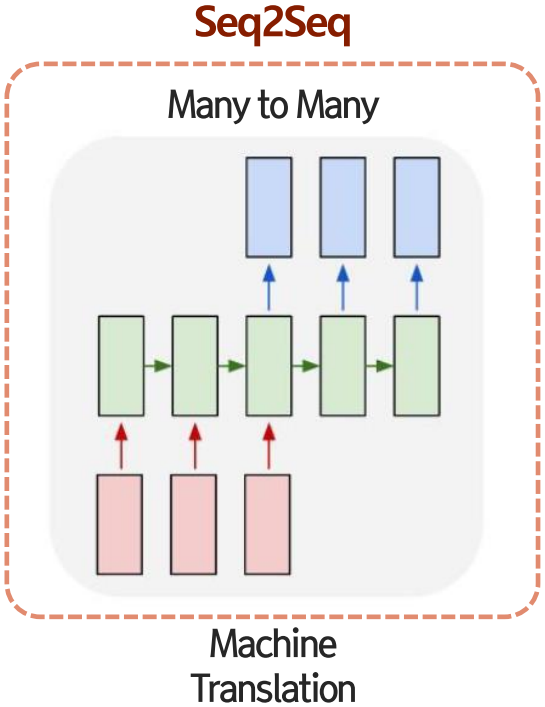
❖ RNN Architecture

- RNN: 연속적인 데이터에서 연속된 변수들 간의 Dependency 를 반영
- 순차적으로 입력하고, 순차적으로 예측하는 알고리즘









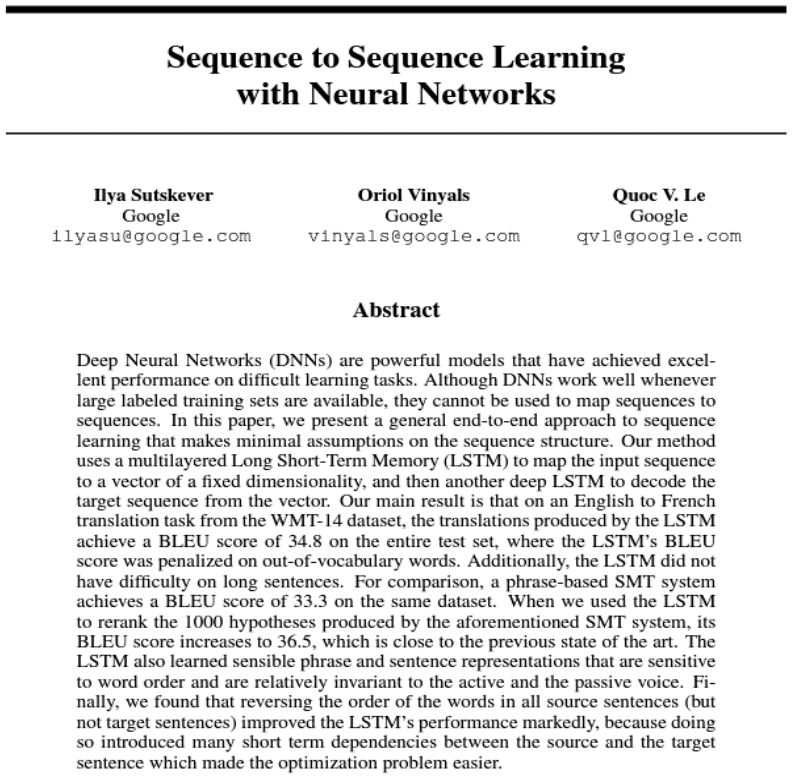
[2] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. Advances in neural information processing systems, 27.

From Transformer to Large Language Model

Seq2Seq

❖ Sequence to Sequence Learning with Neural Networks (2014, NeurIPS)

- 문장을 문장으로 변환하는 모델
- Encoder 와 Decoder 로 이루어져 있으며, 문장 간의 의미적 관계를 학습하여 복잡한 NLP 작업 수행



Sequence to Sequence Learning with Neural Networks

Ilya Sutskever
Google
ilyasu@google.com

Oriol Vinyals
Google
vinyals@google.com

Quoc V. Le
Google
qvl@google.com

Abstract

Deep Neural Networks (DNNs) are powerful models that have achieved excellent performance on difficult learning tasks. Although DNNs work well whenever large labeled training sets are available, they cannot be used to map sequences to sequences. In this paper, we present a general end-to-end approach to sequence learning that makes minimal assumptions on the sequence structure. Our method uses a multilayered Long Short-Term Memory (LSTM) to map the input sequence to a vector of a fixed dimensionality, and then another deep LSTM to decode the target sequence from the vector. Our main result is that on an English to French translation task from the WMT-14 dataset, the translations produced by the LSTM achieve a BLEU score of 34.8 on the entire test set, where the LSTM's BLEU score was penalized on out-of-vocabulary words. Additionally, the LSTM did not have difficulty on long sentences. For comparison, a phrase-based SMT system achieves a BLEU score of 33.3 on the same dataset. When we used the LSTM to rerank the 1000 hypotheses produced by the aforementioned SMT system, its BLEU score increases to 36.5, which is close to the previous state of the art. The LSTM also learned sensible phrase and sentence representations that are sensitive to word order and are relatively invariant to the active and the passive voice. Finally, we found that reversing the order of the words in all source sentences (but not target sentences) improved the LSTM's performance markedly, because doing so introduced many short term dependencies between the source and the target sentence which made the optimization problem easier.

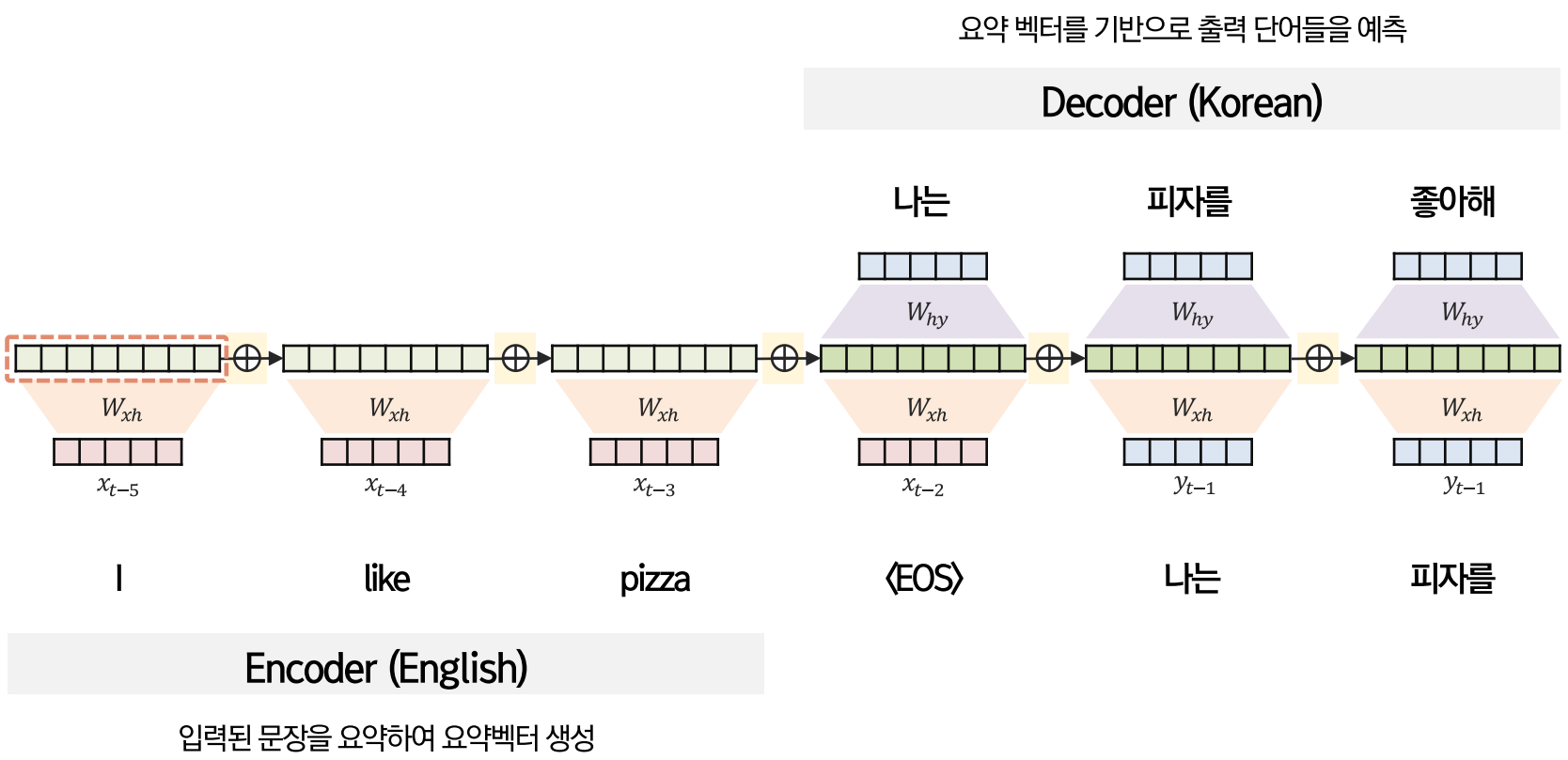
[2] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. Advances in neural information processing systems, 27.

From Transformer to Large Language Model

Seq2Seq

❖ Sequence to Sequence Learning with Neural Networks (2014, NeurIPS)

- Encoder : 입력 시퀀스에 포함된 정보들을 부호화 하여 Context Vector 생성
- Decoder : Encoder 로 부터 Context Vector 를 전달받아 복호화 하여 순차적인 출력 시퀀스 생성

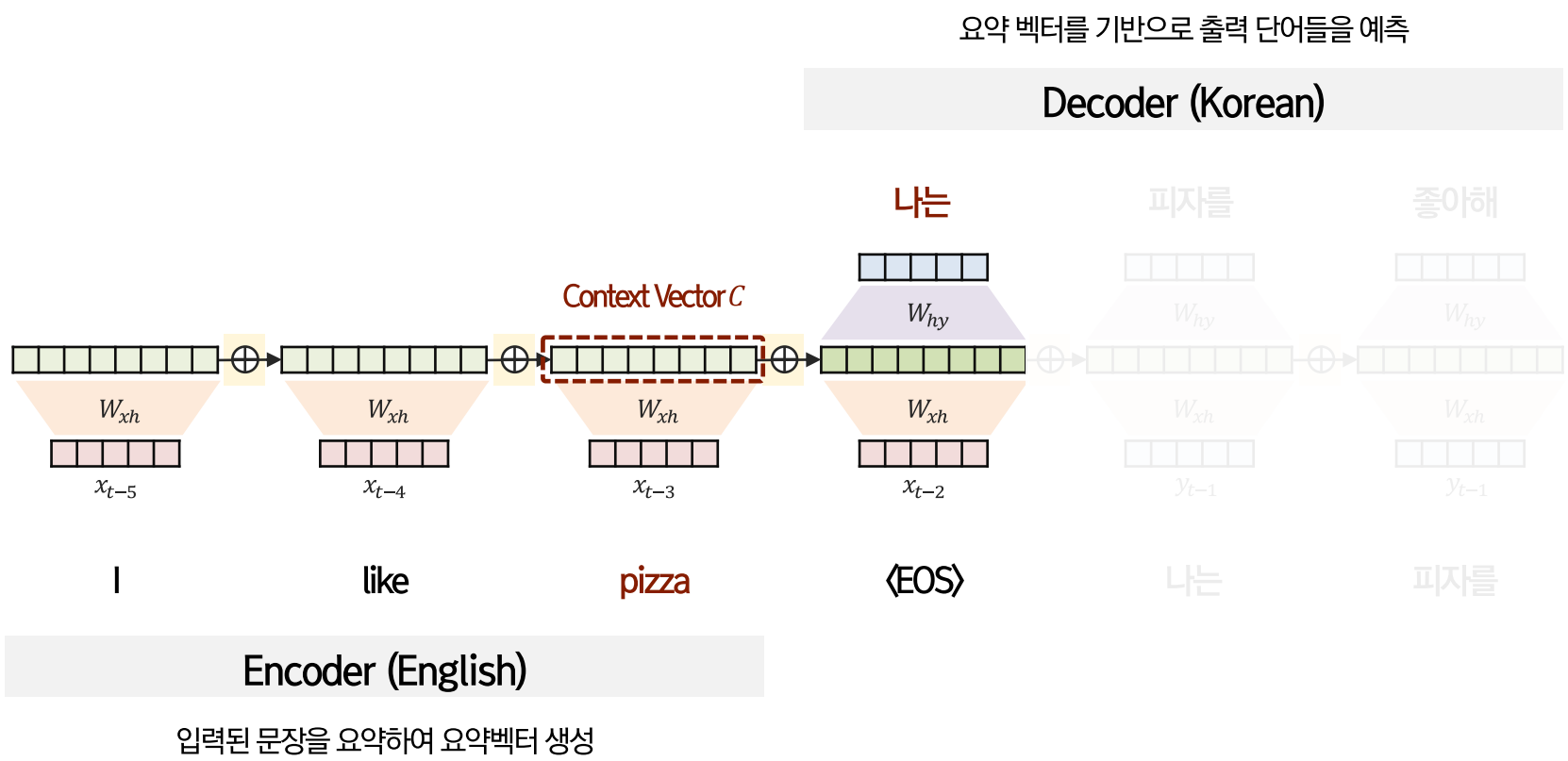


From Transformer to Large Language Model

Seq2Seq

❖ Sequence to Sequence Learning with Neural Networks (2014, NeurIPS)

- Decoder 가 단어 예측 시, Encoder 의 마지막 시점 정보 Context Vector 만을 활용
- 긴 시퀀스를 처리할 때 주어에 해당하는 정보가 희석되는 단점

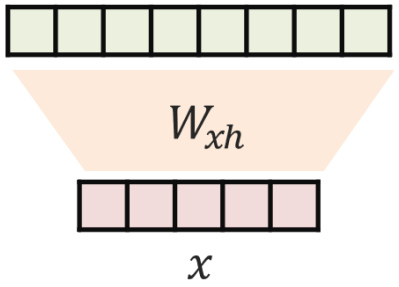


From Transformer to Large Language Model

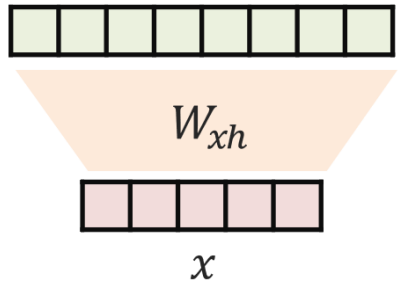
Seq2Seq

❖ Sequence to Sequence Learning with Neural Networks (2014, NeurIPS)

- Decoder 가 단어 예측 시, Encoder 의 마지막 시점 정보 Context Vector 만을 활용
- 긴 시퀀스를 처리할 때 주어에 해당하는 정보가 희석되는 단점
- 필요한 정보를 제한된 길이의 고정벡터에 충분히 표현 불가능



Peter Piper likes pizza



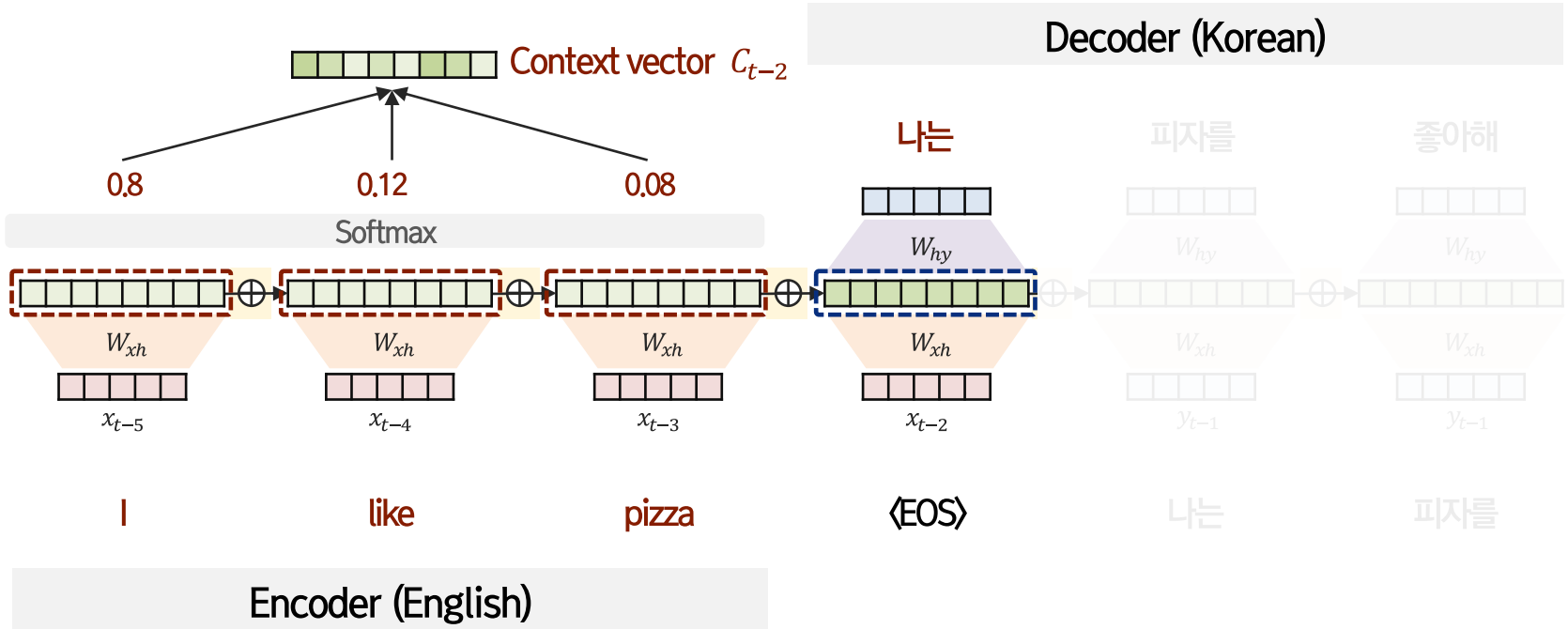
Peter Piper picked a peck of pickled peppers,
A peck of pickled peppers Peter Piper picked;
If Peter Piper picked a peck of pickled peppers, Where's the peck of pickled peppers Peter Piper pickled?

From Transformer to Large Language Model

Attention

❖ Attention in Seq2Seq Learning

- Encoder 의 모든 시점 정보 (Context Vector)를 Decoder 에 전달
- 예측 단어와 관련된 중요한 단어를 위주로 참고
- 입력 문장 길이에 비례하는 정보로 Encoding 하여 정보 손실 발생 방지



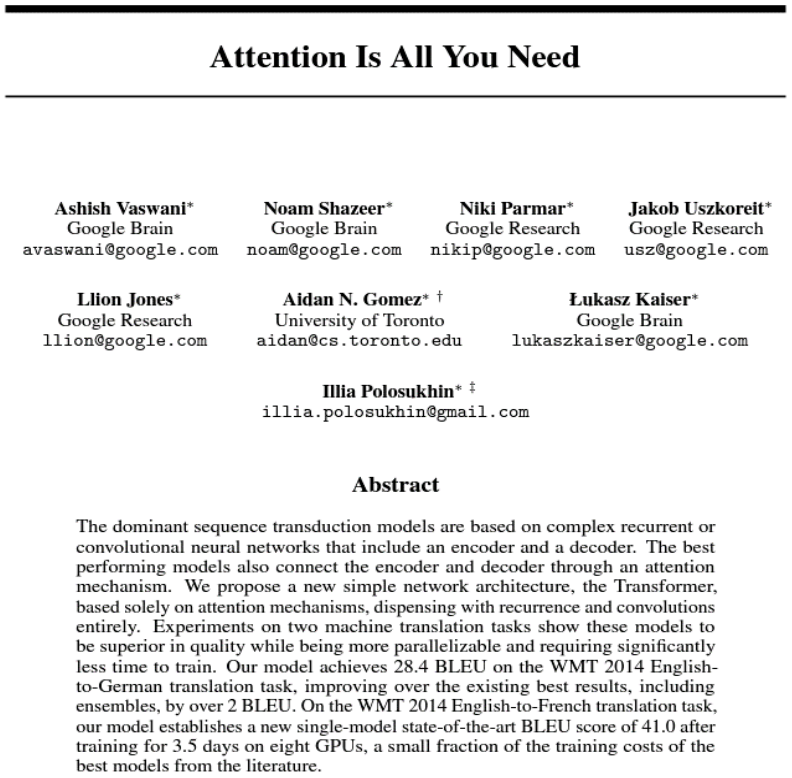
[3] Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

From Transformer to Large Language Model

Transformer

❖ Attention Is All You Need (2017, NeurIPS)

- Seq2Seq 모델 Attention 의 병렬적 사용을 극대화한 모델
- NLP 분야에서 시작되어 다양한 분야에서 응용



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 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.

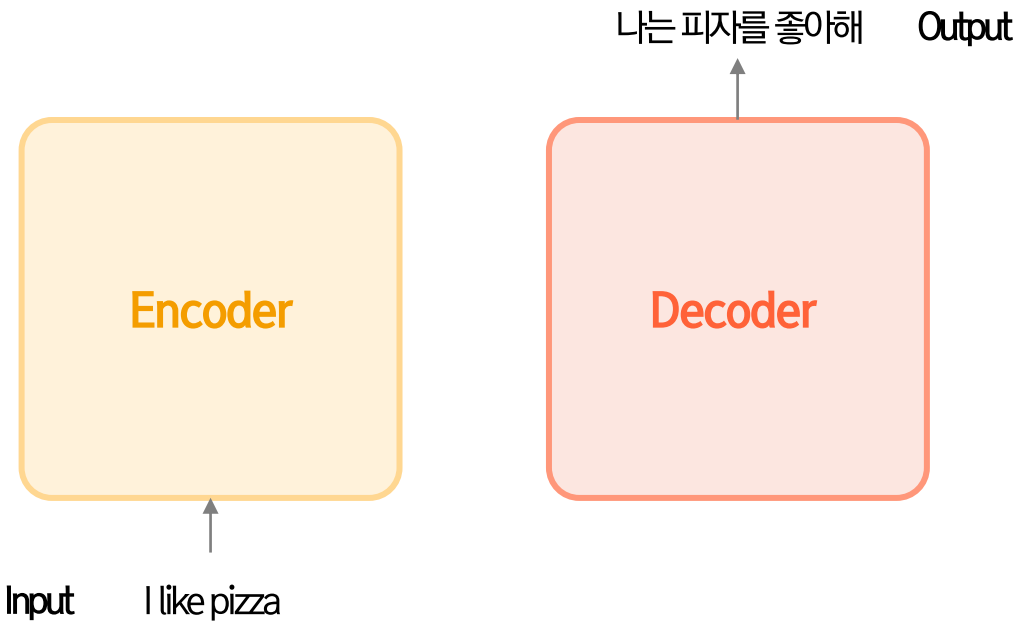
[4] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

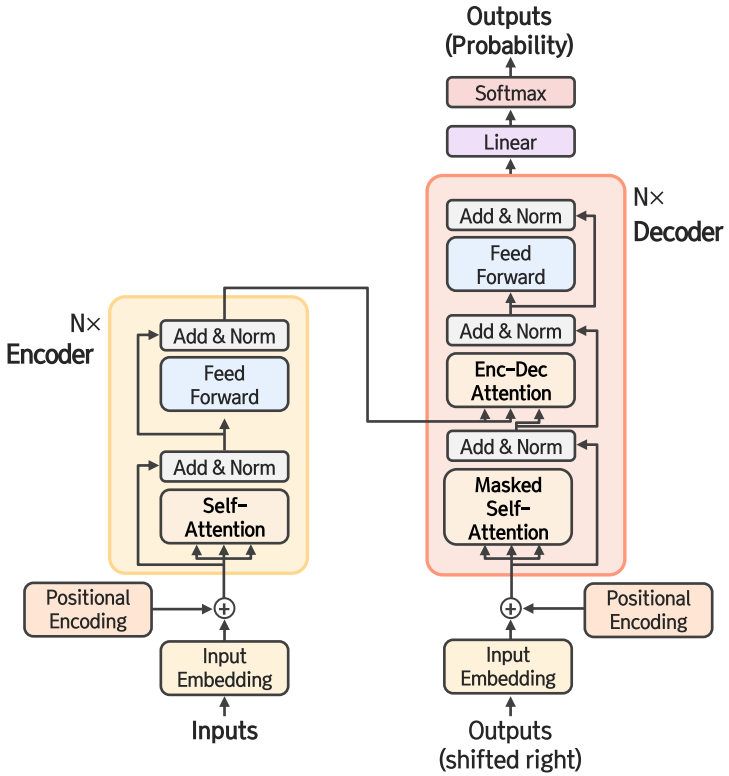
From Transformer to Large Language Model

Transformer

❖ Attention Is All You Need (2017, NeurIPS)

- Encoder-Decoder 구조로 , 각각 동일한 구조의 N개 모듈로 구성
  - ✓ 기본 모델에서 N=6



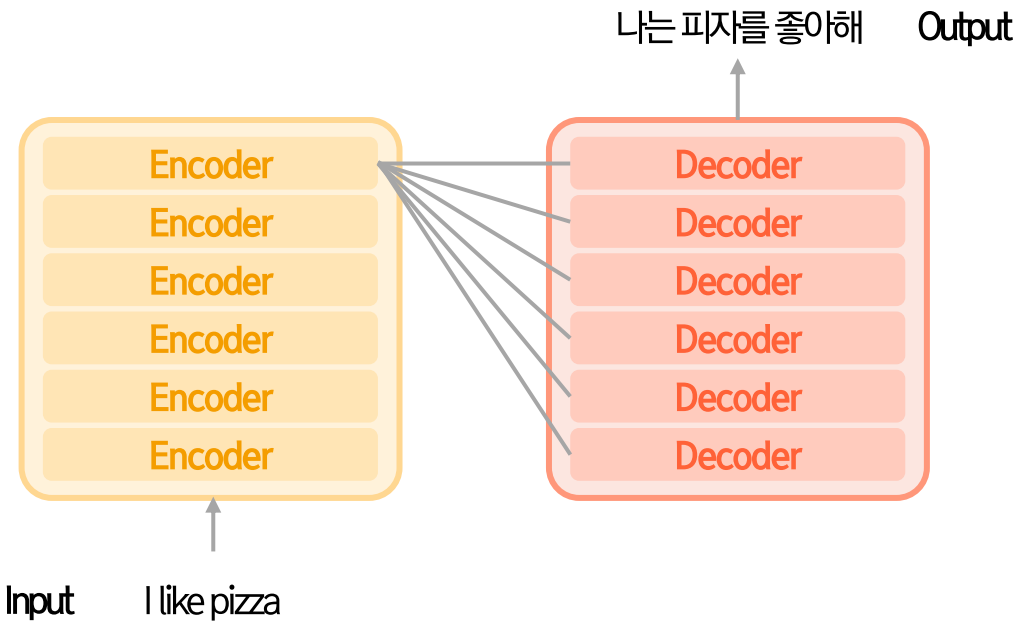


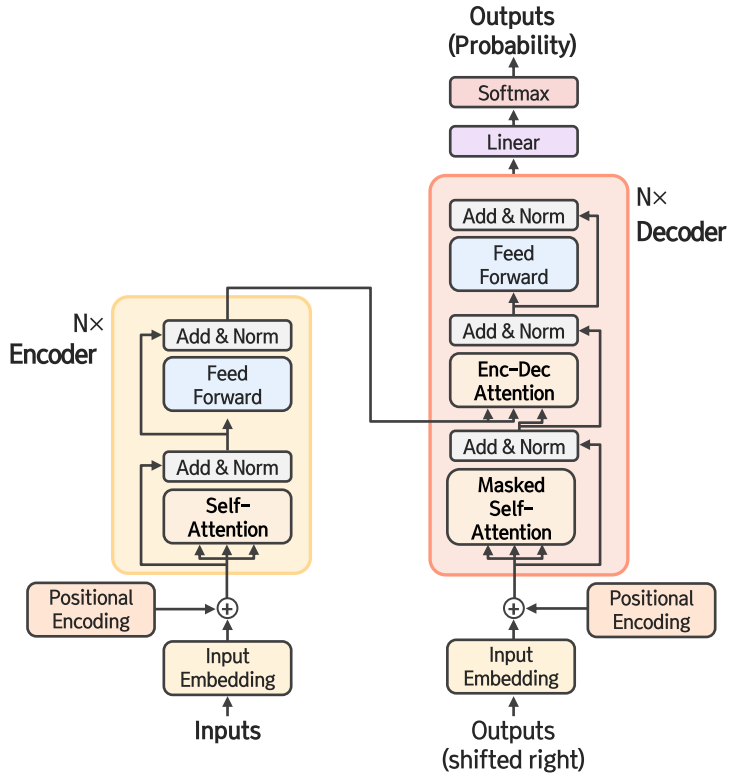
From Transformer to Large Language Model

Transformer

❖ Attention Is All You Need (2017, NeurIPS)

- Encoder-Decoder 구조로 , 각각 동일한 구조의 N개 모듈로 구성
  - ✓ 기본 모델에서 N=6
- 마지막 Encoder 가 모든 단계의 Decoder 에 관여





From Transformer to Large Language Model

Transformer

❖ Attention Is All You Need (2017, NeurIPS)

- Encoder-Decoder 구조로 , 각각 동일한 구조의 N개 모듈로 구성
  - ✓ 기본 모델에서 N=6
- 마지막 Encoder 가 모든 단계의 Decoder 에 관여

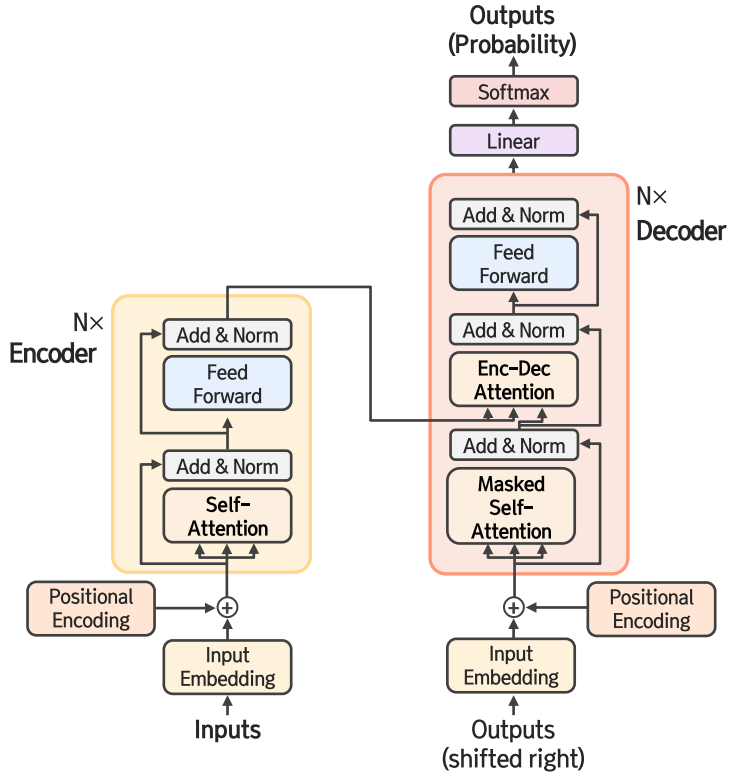
① Embedding

② Positional Encoding

③ Encoder
- Self-Attention
- Feed Forward

④ Decoder
- Masked Self-Attention
- Encoder-Decoder Attention
- Feed Forward

⑤ Prediction



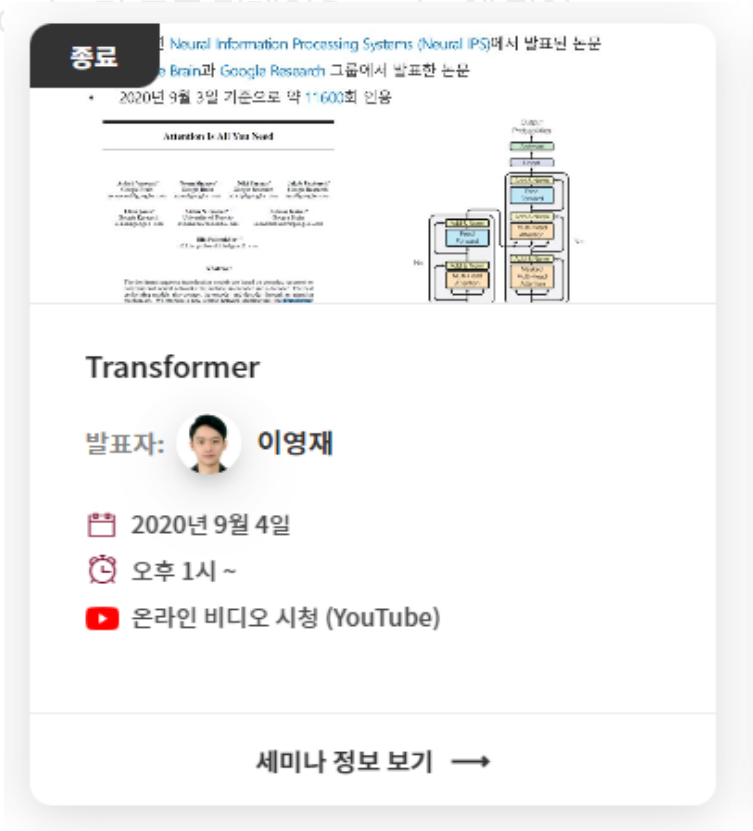
From Transformer to Large Language Model

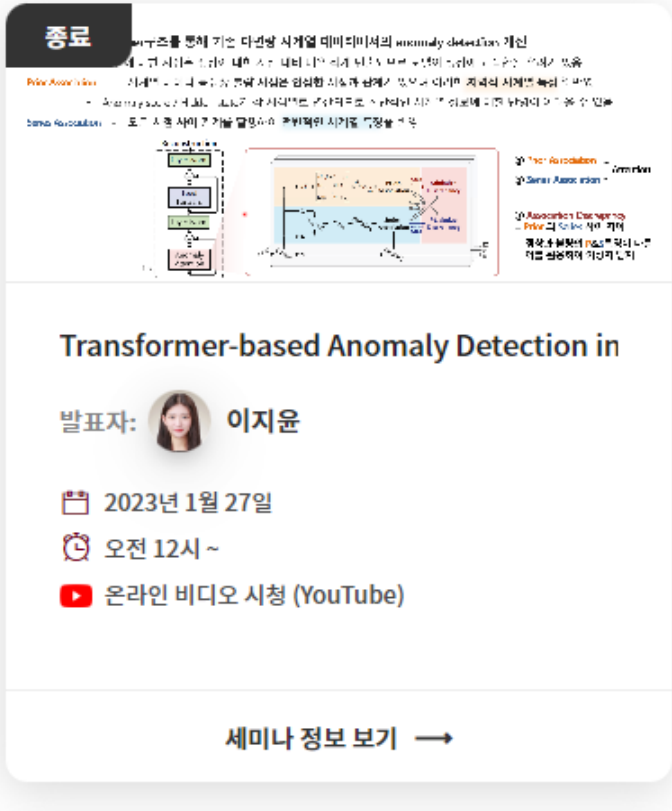
Transformer

❖ Attention Is All You Need (2017, NeurIPS)

- Encoder-Decoder 구조로, 각각 동일한 구조의 N개 모듈로 구성
  - ✓ 기본 모델에서 N=6
- 마지막 Enc

관련 DMQA Open Seminar





From Transformer to Large Language Model

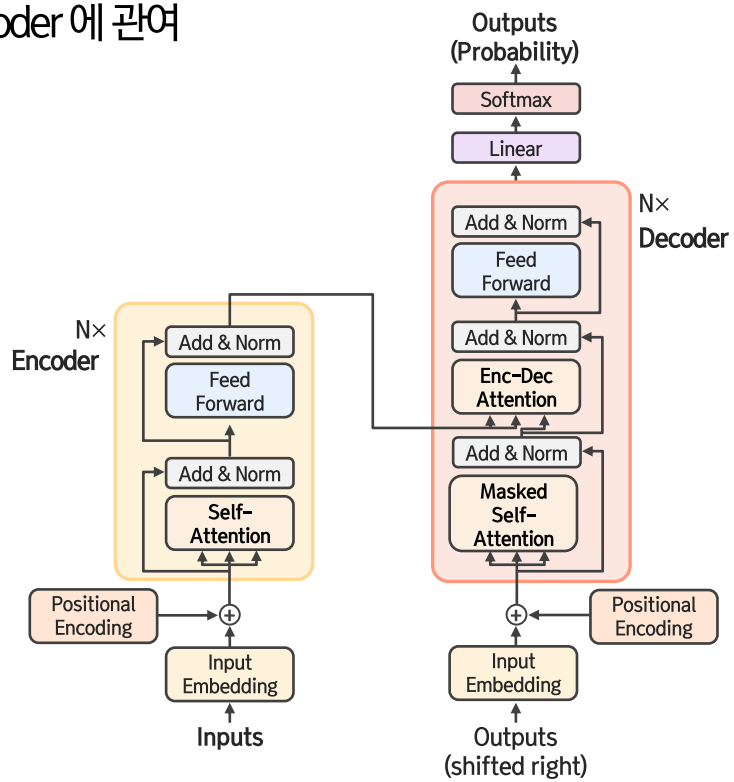
Transformer

❖ Attention Is All You Need (2017, NeurIPS)

- Encoder-Decoder 구조로 , 각각 동일한 구조의 N개 모듈로 구성
  - ✓ 기본 모델에서 N=6
- 마지막 Encoder 가 모든 단계의 Decoder 에 관여

Encoder

Google



Decoder

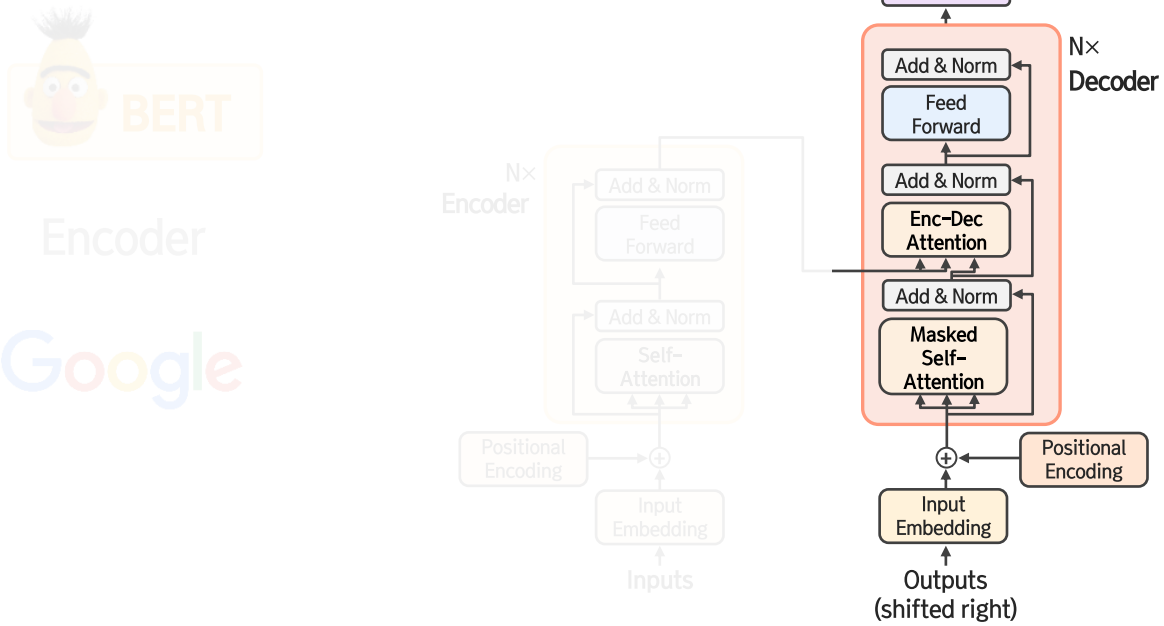
OpenAI

From Transformer to Large Language Model

Transformer

❖ Attention Is All You Need (2017, NeurIPS)

- Encoder-Decoder 구조로 , 각각 동일한 구조의 N개 모듈로 구성
  - ✓ 기본 모델에서 N=6
- 마지막 Encoder 가 모든 단계의 Decoder 에 관여



BERT

Encoder

Google

GPT

Decoder

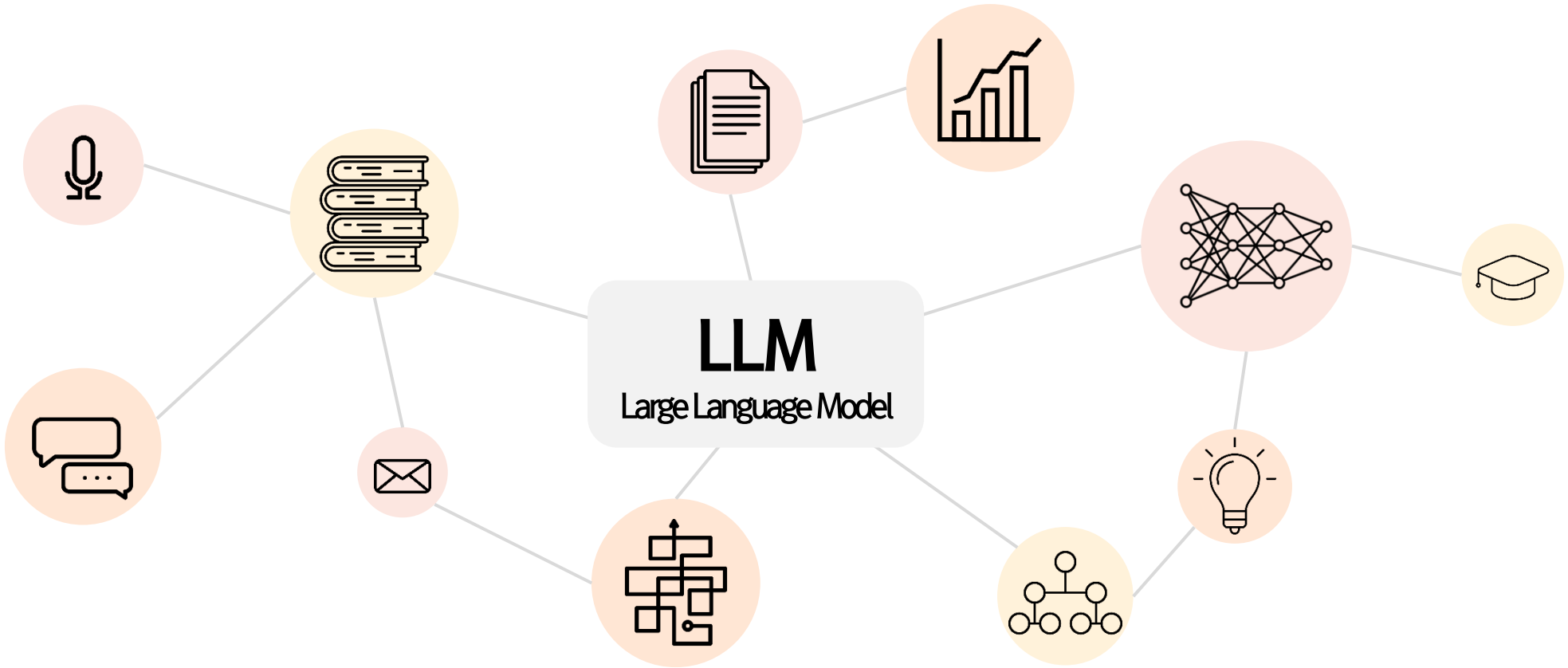
OpenAI

3. Large Language Model (LLM)

Large Language Model (LLM)

❖ Background for LLMs

- LLM : 방대한 텍스트 데이터로 학습된 수천 억 개의 파라미터를 가진 Transformer Language Model
- 언어를 이해하고 복잡한 과업을 수행하는데 강력한 능력



Large Language Model (LLM)

❖ Background for LLMs

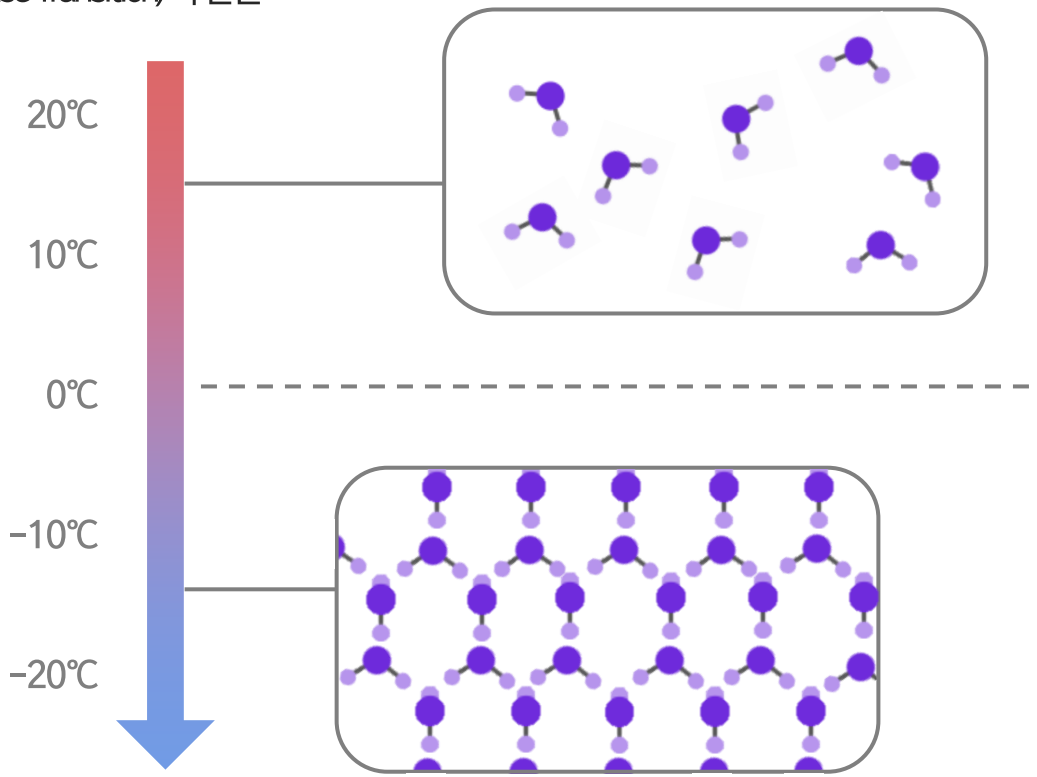
- LLM : 방대한 텍스트 데이터로 학습된 수천 억 개의 파라미터를 가진 Transformer Language Model
- 언어를 이해하고 복잡한 과업을 수행하는데 강력한 능력

Emergent Abilities : 기존 PLM과 구별되는 LLM의 가장 큰 특징

Large Language Model (LLM)

❖ Emergent Abilities for LLMs

- Emergent Abilities : 작은 모델에서는 나타나지 않지만, 거대한 모델에서 발현되는 능력
- 모델의 규모가 일정 수준에 도달하면 성능이 매우 크게 상승
  - ✓ 물리학의 상 전이 (Phase Transition) 와 관련



Large Language Model (LLM)

❖ Emergent Abilities for LLMs

- Emergent Abilities : 작은 모델에서는 나타나지 않지만, 거대한 모델에서 발현되는 능력
- LLM 에 대한 대표적인 Emergent Abilities : In-context Learning, Instruction Following, Step-by-step Reasoning

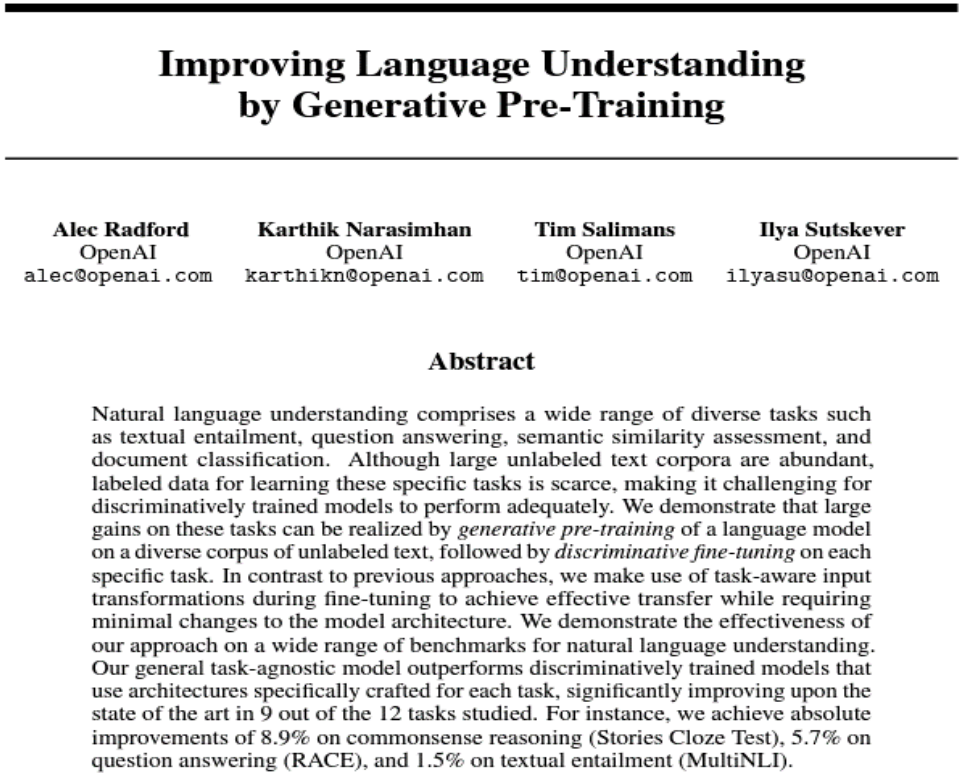
In-context Learning	언어모델에 자연어가 제공 되었다고 가정 GPT Series 모델 중 GPT-1, GPT-2 에 비해 GPT-3 가 가장 강력한 성능
Instruction Following	Instruction Tuning을 통해 학습하지 않았던 Task 에 대해서도 잘 수행됨 모델 크기가 커지면 튜닝된 모델이 훨씬 뛰어난 능력
Step-by-step Reasoning	작은 모델의 경우 여러 추론 단계가 필요한 복잡한 작업 어려움 특정 전략을 통해 모델 크기가 클수록 높은 성능 향상

Large Language Model (LLM)

GPT Series : GPT-1

❖ Improving Language Understanding by Generative Pre-Training (2018)

- 비지도 사전학습과 지도 미세조정을 조합한 NLP에 대한 '준지도학습' 제시
- 다양한 작업에 쉽게 전이가능한 보편적인 표현 학습
  - ✓ 최소한의 모델 구조 변화를 통해 목적에 맞게 전이학습 가능



Improving Language Understanding by Generative Pre-Training

Alec Radford OpenAI alec@openai.com
Karthik Narasimhan OpenAI karthikn@openai.com
Tim Salimans OpenAI tim@openai.com
Ilya Sutskever OpenAI ilyasu@openai.com

Abstract

Natural language understanding comprises a wide range of diverse tasks such as textual entailment, question answering, semantic similarity assessment, and document classification. Although large unlabeled text corpora are abundant, labeled data for learning these specific tasks is scarce, making it challenging for discriminatively trained models to perform adequately. We demonstrate that large gains on these tasks can be realized by *generative pre-training* of a language model on a diverse corpus of unlabeled text, followed by *discriminative fine-tuning* on each specific task. In contrast to previous approaches, we make use of task-aware input transformations during fine-tuning to achieve effective transfer while requiring minimal changes to the model architecture. We demonstrate the effectiveness of our approach on a wide range of benchmarks for natural language understanding. Our general task-agnostic model outperforms discriminatively trained models that use architectures specifically crafted for each task, significantly improving upon the state of the art in 9 out of the 12 tasks studied. For instance, we achieve absolute improvements of 8.9% on commonsense reasoning (Stories Cloze Test), 5.7% on question answering (RACE), and 1.5% on textual entailment (MultiNLI).

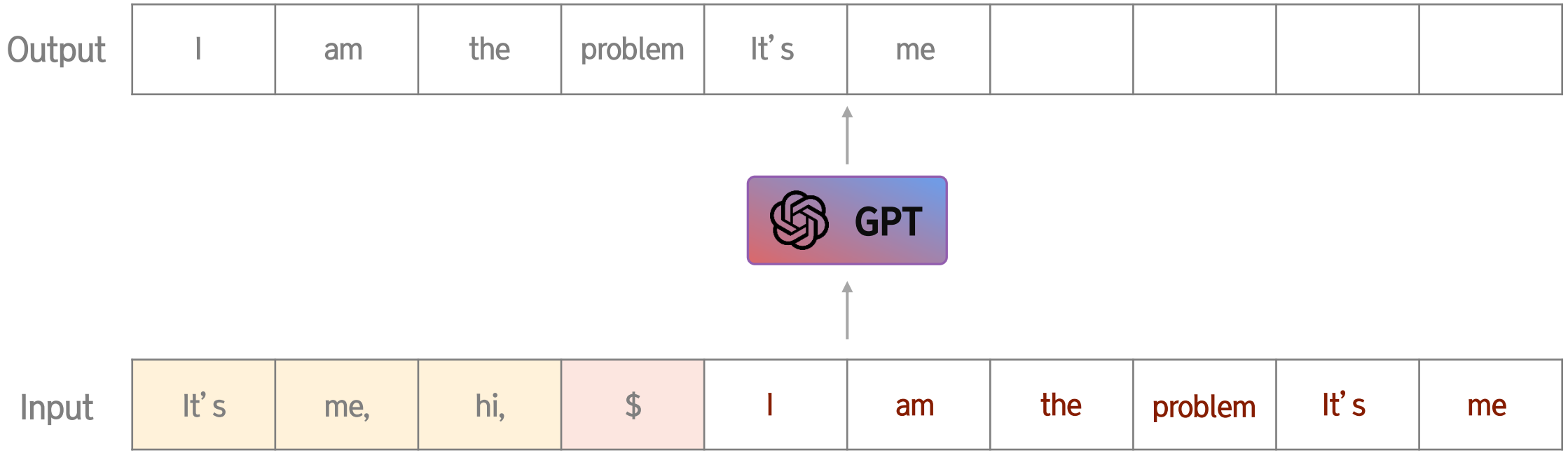
[5] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

Large Language Model (LLM)

GPT Series : GPT-1

❖ GPT : Generative Pre-trained Transformer

- Transformer 의 생성형 Decoder 구조 기반의 Autoregressive 모델
- Autoregressive : 주어진 입력의 일부를 사용하여 이후에 오는 부분을 예측
  - ✓ 주어진 연속된 단어의 다음 단어를 맞추는 방식

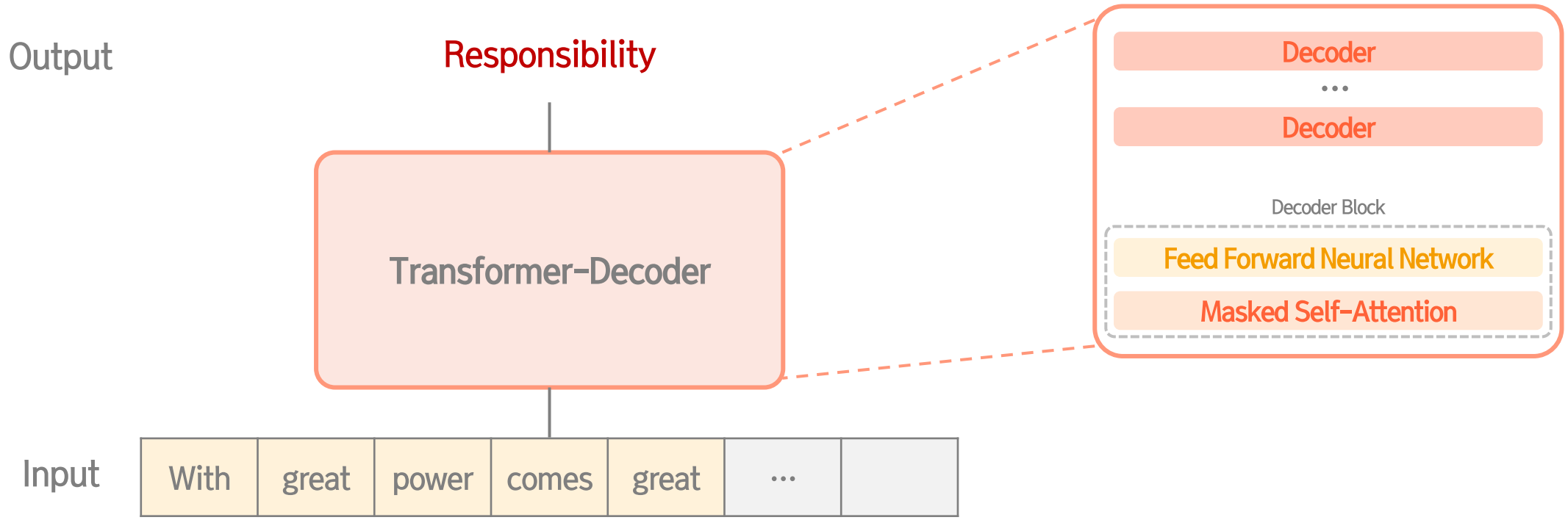


Large Language Model (LLM)

GPT Series : GPT-1

❖ GPT : Generative Pre-trained Transformer

- Transformer 의 생성형 Decoder 구조 기반의 Autoregressive 모델
- Autoregressive : 주어진 입력의 일부를 사용하여 이후에 오는 부분을 예측
  - ✓ 주어진 연속된 단어의 다음 단어를 맞추는 방식



Large Language Model (LLM)

GPT Series : GPT-1

❖ GPT Framework

- Step 1 : 대량의 말뭉치로 대용량의 언어모델 학습 (Unsupervised Pre-training)
  - ✓ 지금까지의 토큰 순서를 기반으로 다음 토큰이 나타날 확률 추정

Objective Function

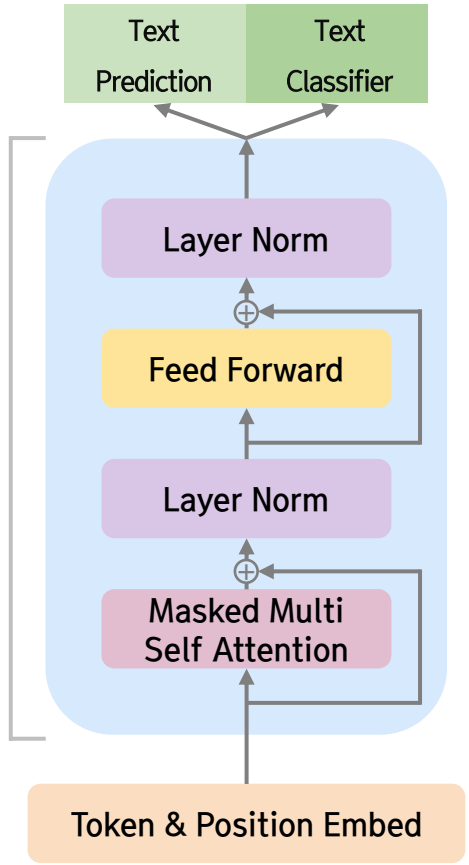
Neural Network Parameters

$$L_1(u) = \sum_i \log P(u_i | u_{i-k}, \ldots, u_{i-1}; \theta)$$

Conditional Probability

u : $\{u_1, u_2, \ldots, y_n\}$ Token Set

k : Window Size



Large Language Model (LLM)

GPT Series : GPT-1

❖ GPT Framework

- Step 1 : 대량의 말뭉치로 대용량의 언어모델 학습 (Unsupervised Pre-training)
  - ✓ Text & Position Embed : Context Vector U

Dataset

Bob likes pizza

Bob hates beans

Bob	0
likes	1
hates	2
pizza	3
beans	4

Dataset

0 1 3

0 2 4

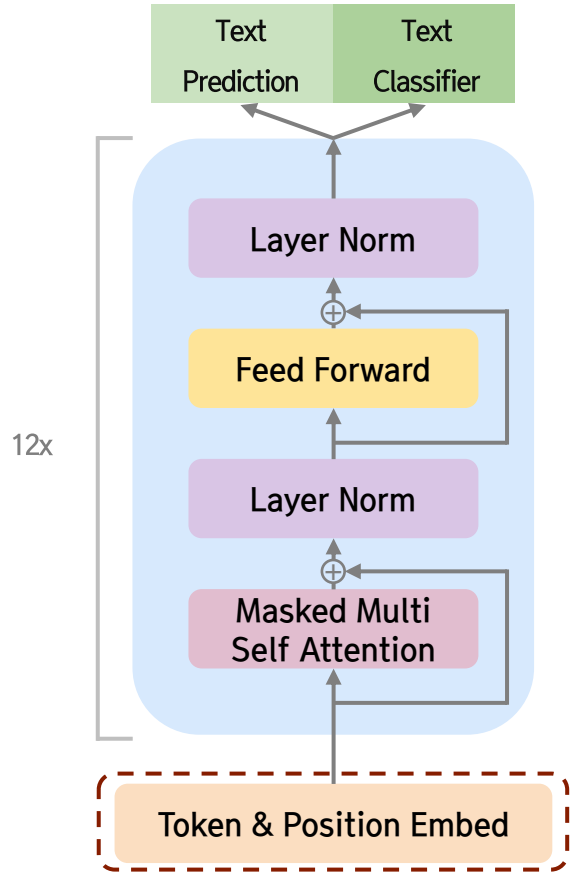
One-Hot Encoding

[[1 0 0 0 0],[0 1 0 0 0],[0 0 0 1 0]]

[[1 0 0 0 0],[0 0 1 0 0],[0 0 0 0 1]]

Context Vector U

[2,3,5]

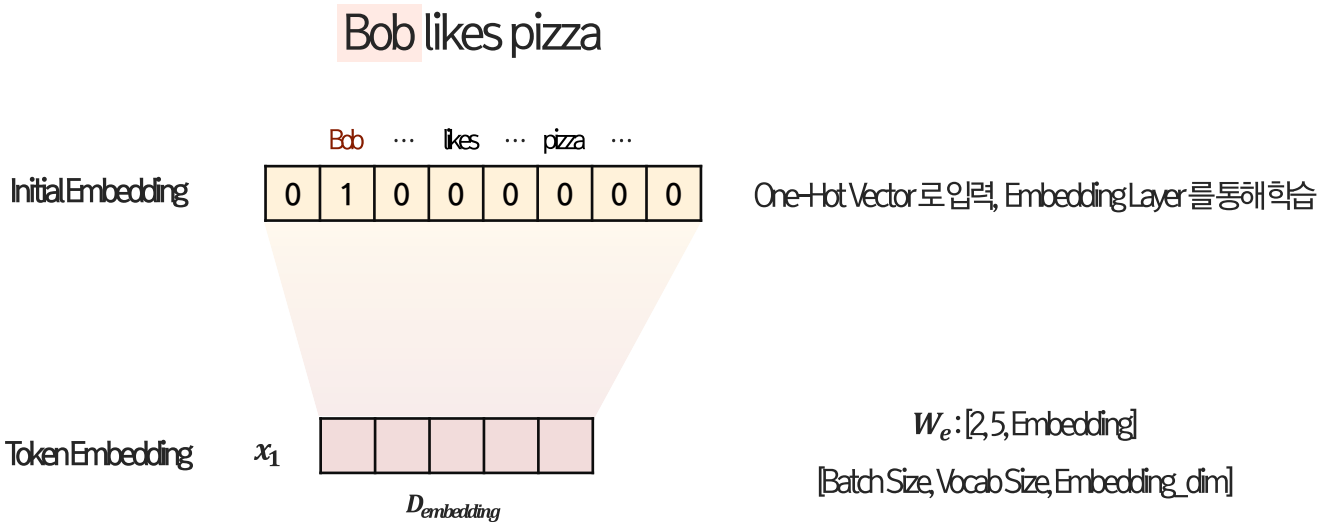


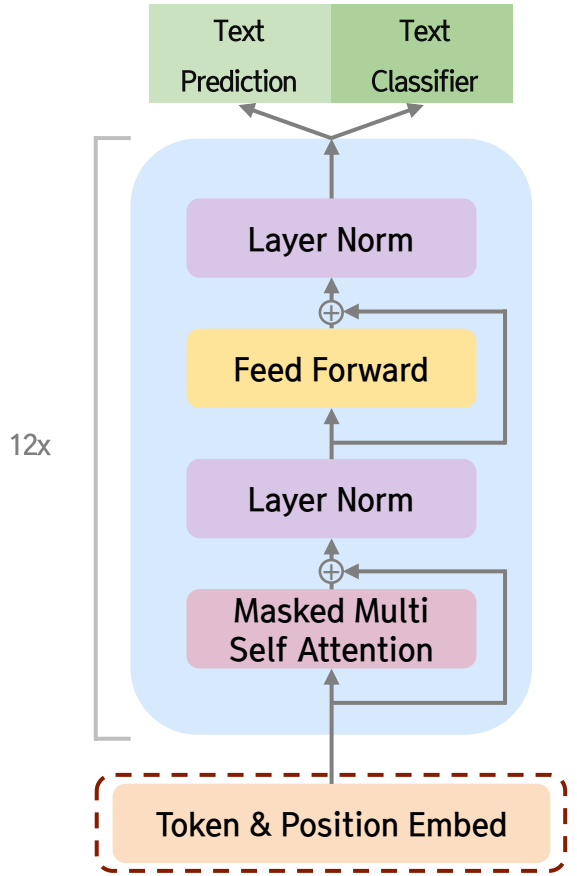
Large Language Model (LLM)

GPT Series : GPT-1

❖ GPT Framework

- Step 1 : 대량의 말뭉치로 대용량의 언어모델 학습 (Unsupervised Pre-training)
  - ✓ Text & Position Embed : Token Embedding Matrix W_e
  - ✓ 유사한 단어는 유사한 값을 가지도록 Embedding





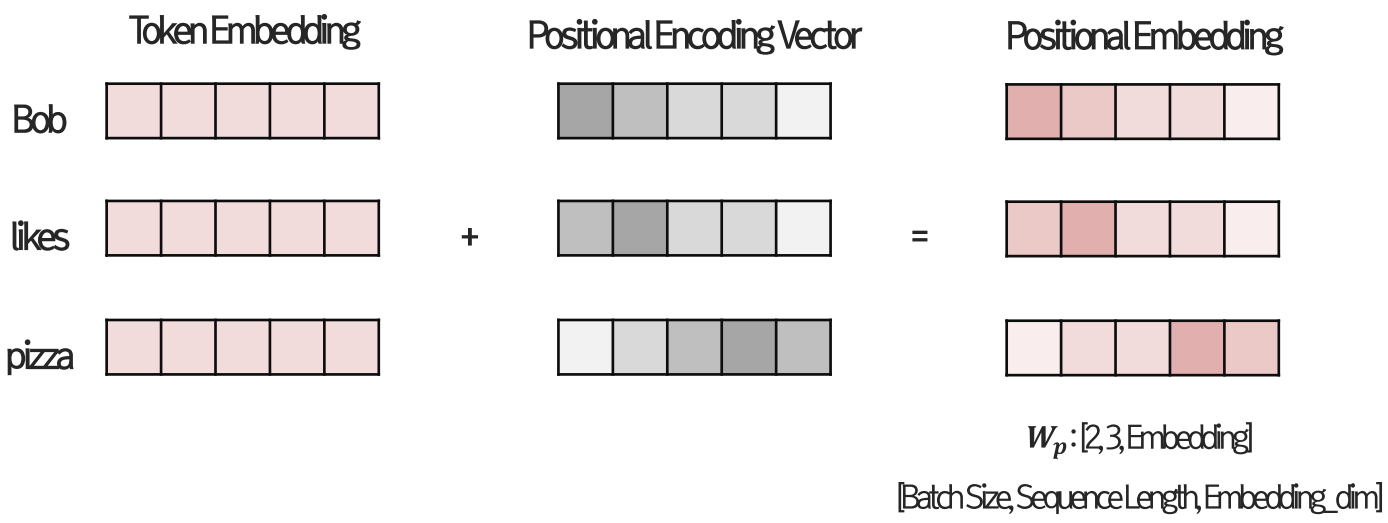
https://medium.com/@hari4om/word-embedding-d816f643140

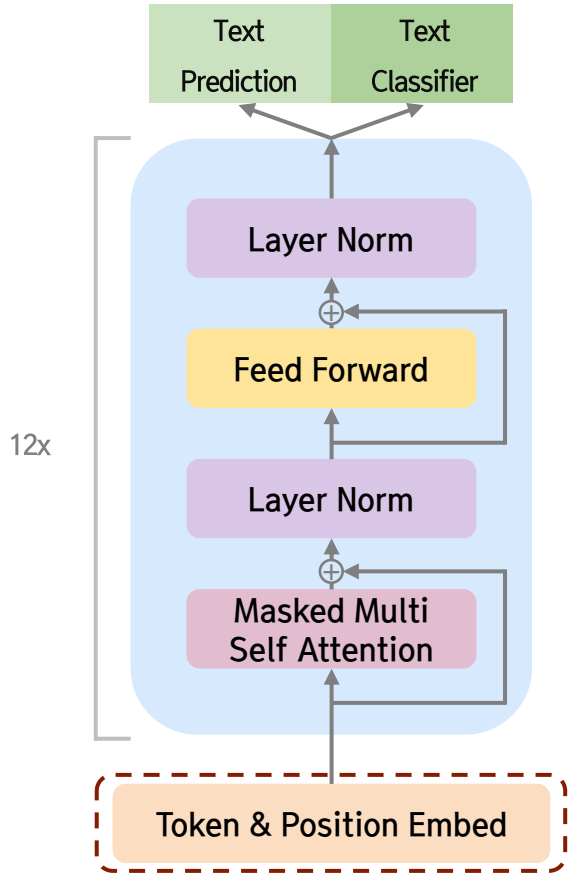
Large Language Model (LLM)

GPT Series : GPT-1

❖ GPT Framework

- Step 1 : 대량의 말뭉치로 대용량의 언어모델 학습 (Unsupervised Pre-training)
  - ✓ Text & Position Embed : Position Embedding Matrix W_p
  - ✓ 단어 사이의 순차성 부여를 위해 -1 ~ 1 범위의 삼각함수 활용 (Positional Encoding)





https://medium.com/@hari4om/word-embedding-d816f643140

Large Language Model (LLM)

GPT Series : GPT-1

❖ GPT Framework

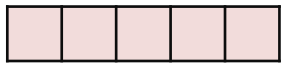
- Step 1 : 대량의 말뭉치로 대용량의 언어모델 학습 (Unsupervised Pre-training)

[[1 0 0 0 0],[0 1 0 0 0],[0 0 0 1 0]]
[[1 0 0 0 0],[0 0 1 0 0],[0 0 0 0 1]]

Context Vector U

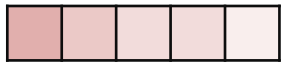
Token Embedding

W_e



Positional Embedding

W_p

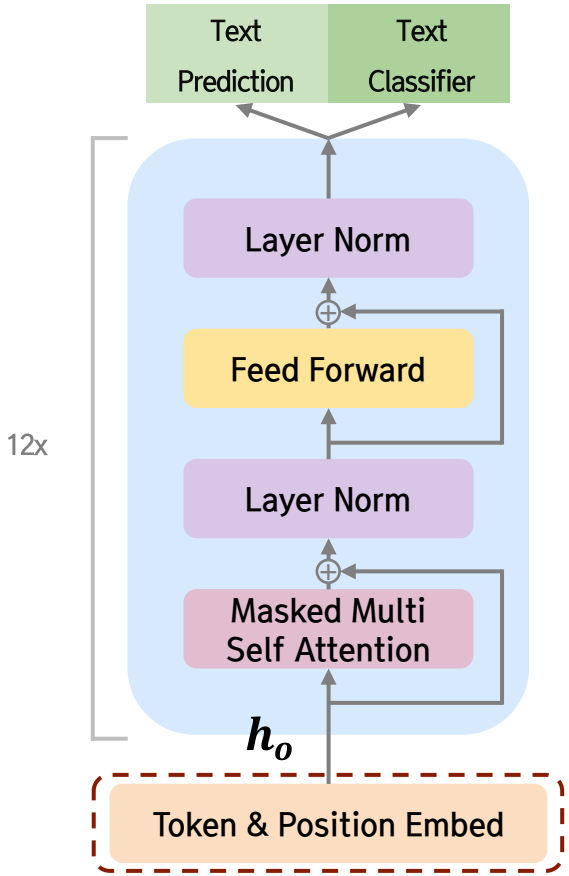


$$h_o = UW_e + W_p$$

U : Context Vector

W_e : Token Embedding

W_p : Position Embedding



Large Language Model (LLM)

GPT Series : GPT-1

❖ GPT Framework

- Step 1 : 대량의 말뭉치로 대용량의 언어모델 학습 (Unsupervised Pre-training)
- Step 2 : Labeled Data를 사용하여 목표 Task에 맞게 미세조정 (Supervised Fine-tuning)

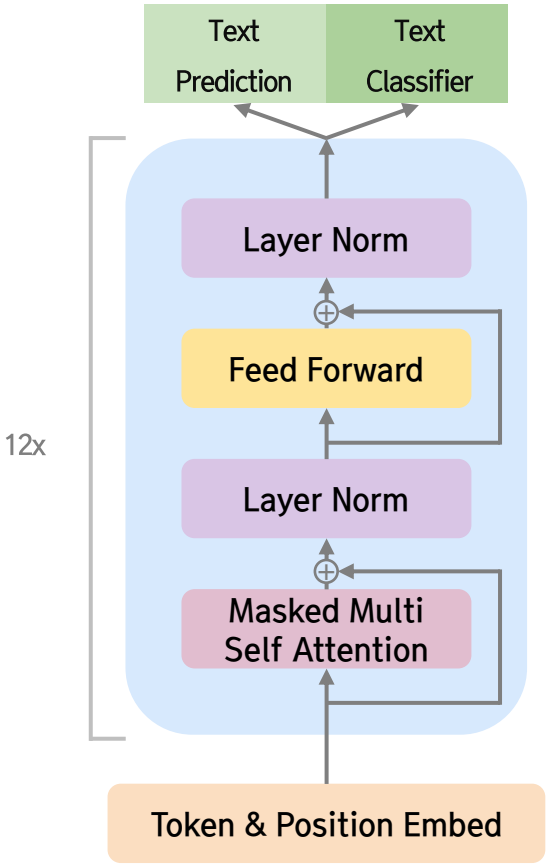
Objective Function

$$P(y|x^1, \ldots, x^m) = softmax(h_l^m W_y)$$

$$L_2(c) = \sum_{(x,y)} \log P(y|x^1, \ldots, x^m)$$

$x^1, \ldots, x^m$: Input Tokens

c : Labeled Dataset



Large Language Model (LLM)

GPT Series : GPT-1

❖ GPT Framework

- Step 1 : 대량의 말뭉치로 대용량의 언어모델 학습 (Unsupervised Pre-training)
- Step 2 : Labeled Data를 사용하여 목표 Task에 맞게 미세조정 (Supervised Fine-tuning)

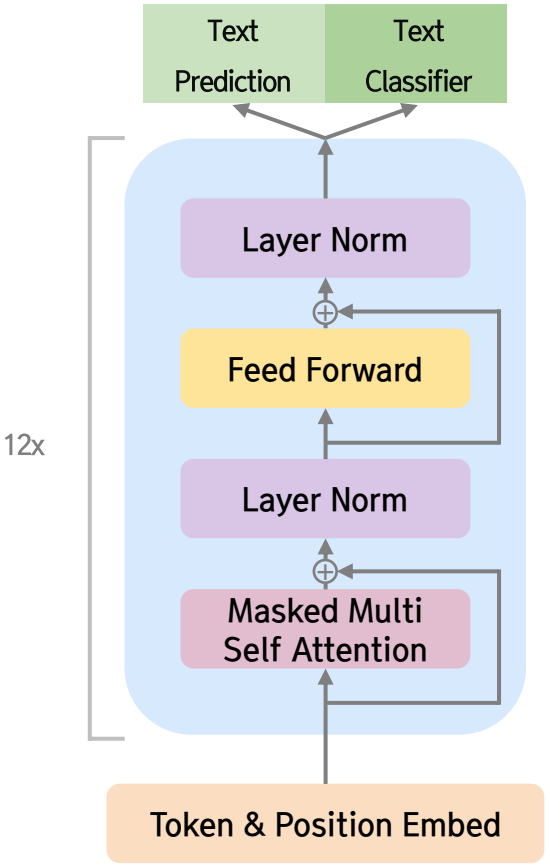
Objective Function

$$L_1(u) = \sum_i \log P(u_i | u_{i-k}, \ldots, u_{i-1}; \theta)$$

$$L_2(c) = \sum_{(x,y)} \log P(y | x^1, \ldots, x^m)$$

새로운 데이터에 대해서도 Step 1 수행

$$L_3(c) = L_2(c) + \lambda L_1(c)$$

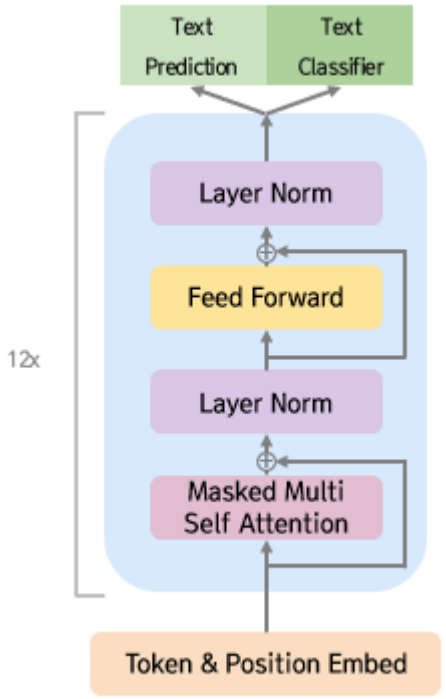


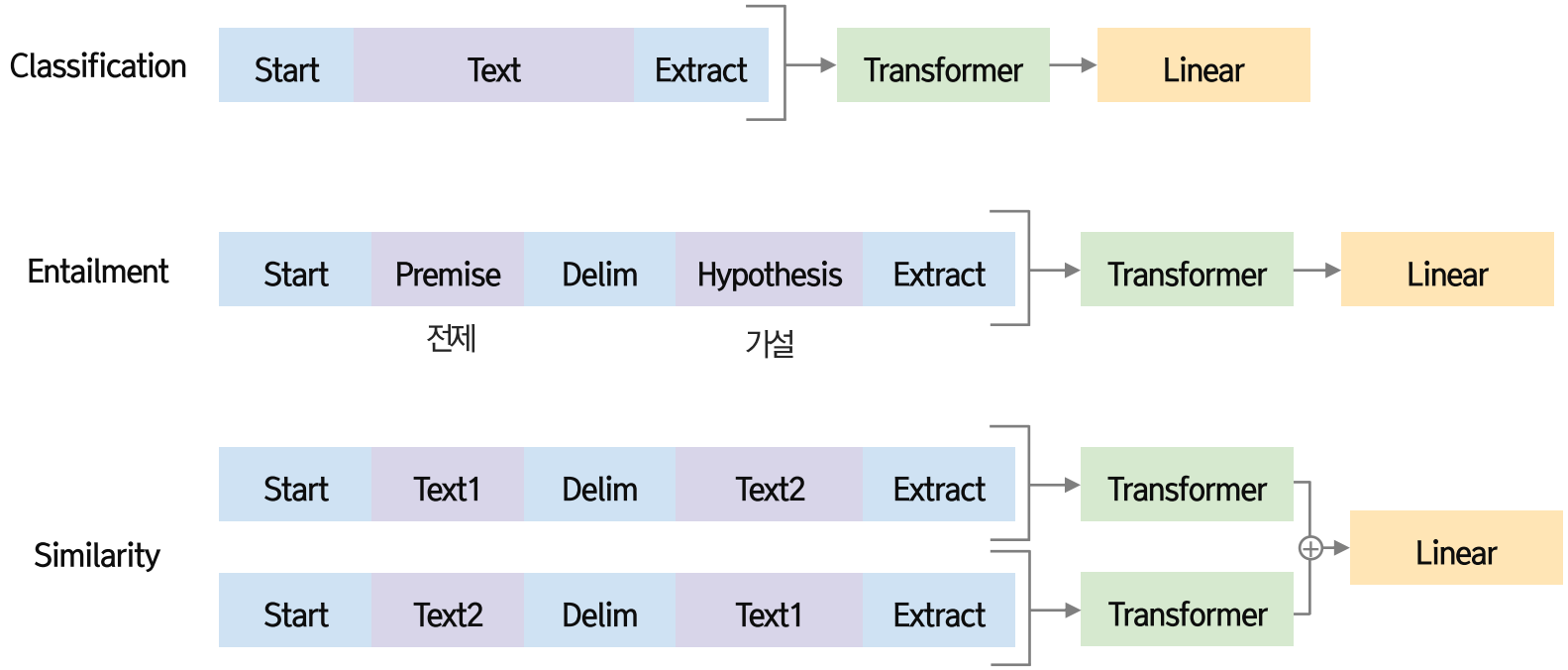
Large Language Model (LLM)

GPT Series : GPT-1

❖ Task-Specific Input Transformations

- Step 2 의 미세조정 과정에서 모델을 최소화로 변환하기 위해 특정 Task에 대한 Input 을 Ordered Sequence 로 변환
  - ✓ Classification : Start, End Token 을 Input Sequence 에 추가
  - ✓ Delimiter Token 을 다른 Example 사이에 추가



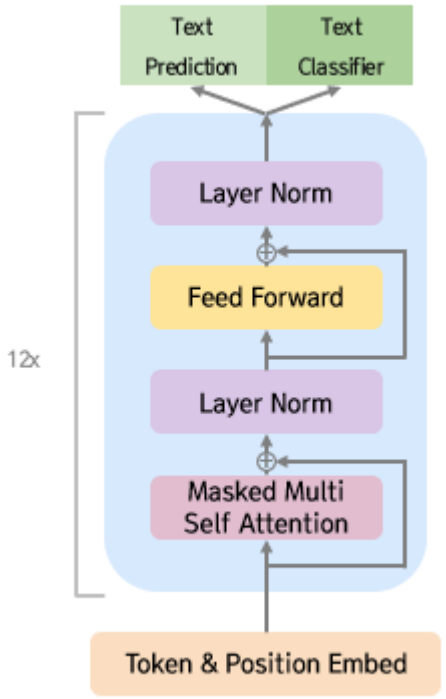


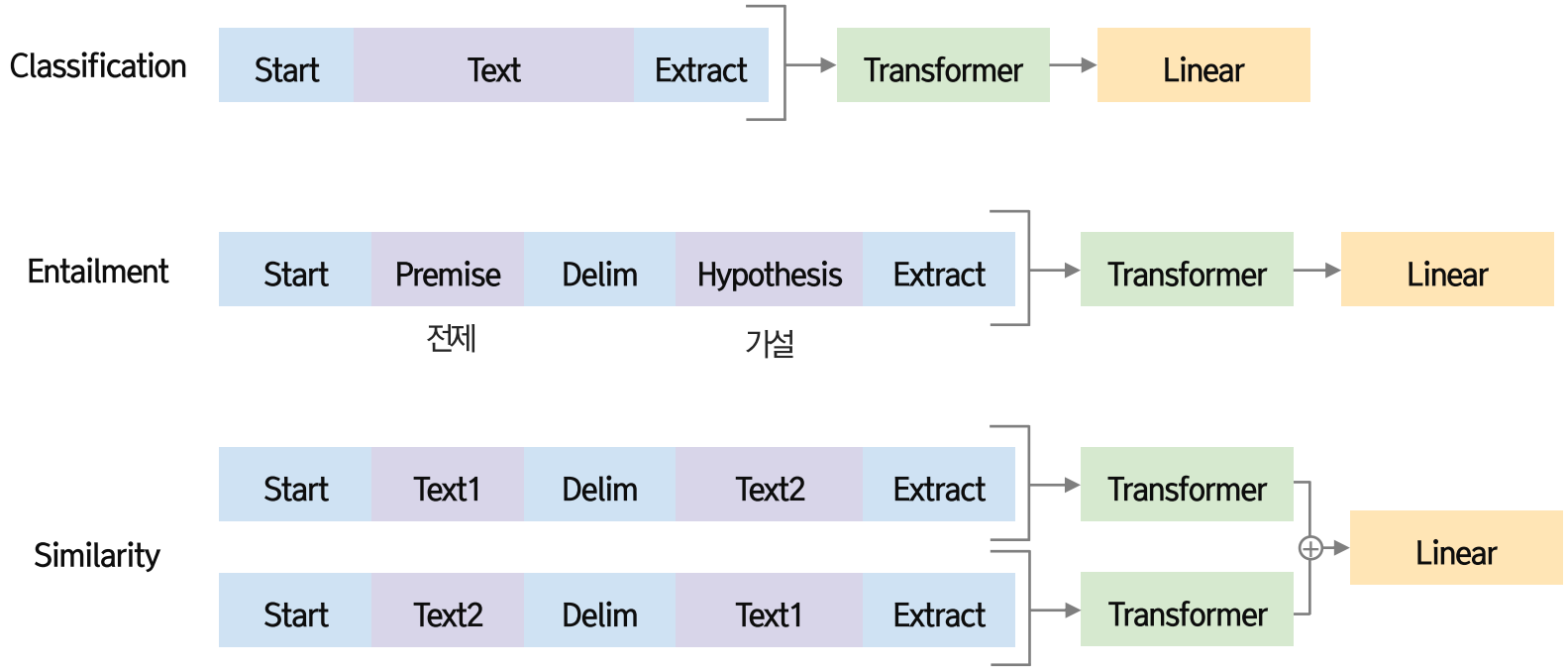
Large Language Model (LLM)

GPT Series : GPT-1

❖ Contribution of GPT-1

- GPT Series 모델의 핵심 구조 설정, 다음 단어 예측을 위한 기본 원칙 확립
- Generative Pre-Training 이 모델의 일반화 성능을 높이는데 효과적임을 확인
- 사전학습을 통해 다양한 NLP Tasks 에서 Zero-shot 성능 향상 입증



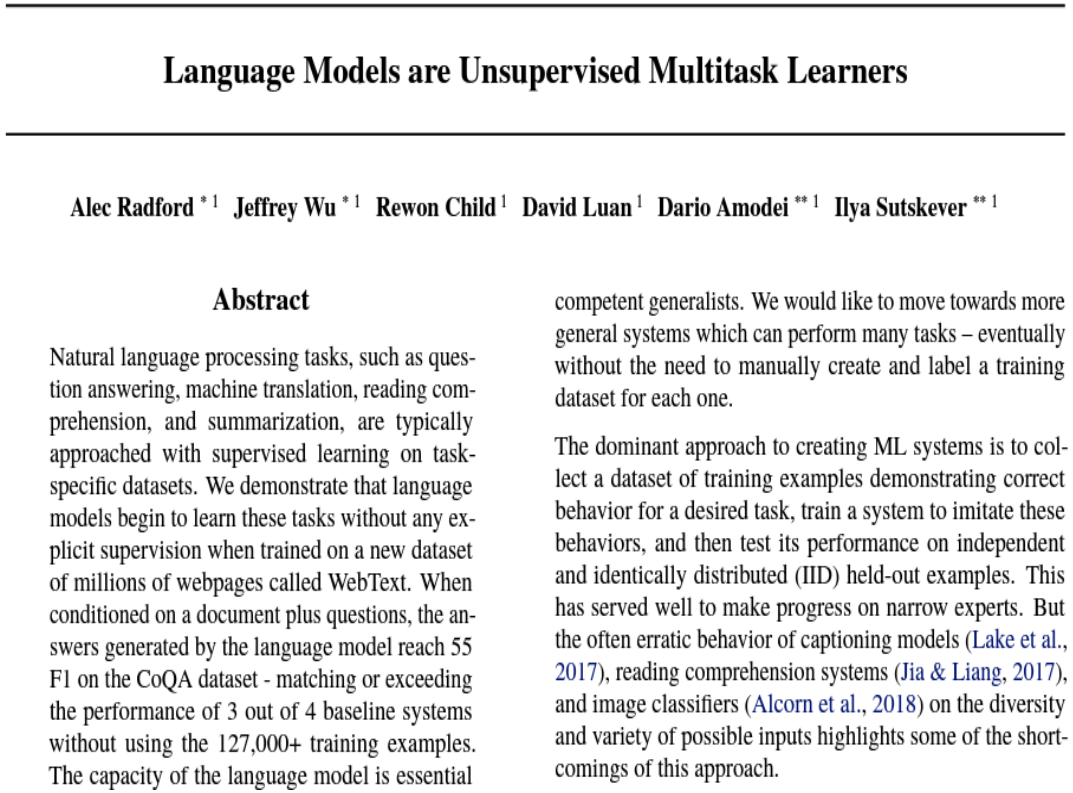


Large Language Model (LLM)

GPT Series : GPT-2

❖ Language Models are Unsupervised Multitask Learners (2019)

- GPT-1 에서 더 큰 데이터셋 & 더 많은 파라미터 학습을 통해 더 강력한 언어모델 학습
- 동일한 Unsupervised Model 을 활용하여 Multiple Tasks 학습
  - ✓ 모델이 주어진 입력 문장들을 통해 적합한 Task 를 유추하고 적절한 답을 내는 것이 목표



Language Models are Unsupervised Multitask Learners

Alec Radford [*1] Jeffrey Wu [*1] Rewon Child [1] David Luan [1] Dario Amodei [**1] Ilya Sutskever [**1]

Abstract

Natural language processing tasks, such as question answering, machine translation, reading comprehension, and summarization, are typically approached with supervised learning on task-specific datasets. We demonstrate that language models begin to learn these tasks without any explicit supervision when trained on a new dataset of millions of webpages called WebText. When conditioned on a document plus questions, the answers generated by the language model reach 55 F1 on the CoQA dataset - matching or exceeding the performance of 3 out of 4 baseline systems without using the 127,000+ training examples. The capacity of the language model is essential to the success of zero-shot task transfer and increasing it improves performance in a log-linear fashion across tasks. Our largest model is a state-of-the-art model on 7 of 8 tested language modeling datasets. The model is also competent generalists. We would like to move towards more general systems which can perform many tasks – eventually without the need to manually create and label a training dataset for each one.

The dominant approach to creating ML systems is to collect a dataset of training examples demonstrating correct behavior for a desired task, train a system to imitate these behaviors, and then test its performance on independent and identically distributed (IID) held-out examples. This has served well to make progress on narrow experts. But the often erratic behavior of captioning models (Lake et al., 2017), reading comprehension systems (Jia & Liang, 2017), and image classifiers (Alcorn et al., 2018) on the diversity and variety of possible inputs highlights some of the shortcomings of this approach.

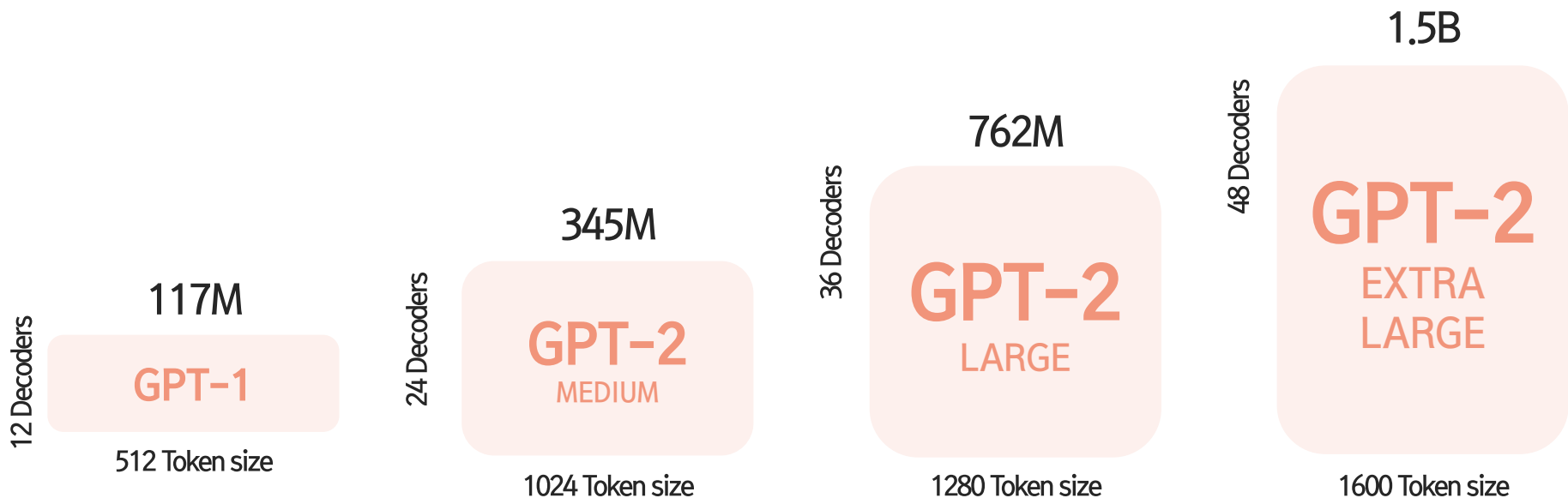
[6] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

Large Language Model (LLM)

GPT Series : GPT-2

❖ GPT-2 Architecture

- GPT-1 에서 더 큰 데이터셋 & 더 많은 파라미터 학습을 통해 더 강력한 언어모델 학습
  - ✓ GPT-1 보다 10배 많은 1.5 Billion 개의 Parameters 사용
- GPT-2 내부 버전에 따라 Parameter와 Layer 의 개수를 다르게 지정



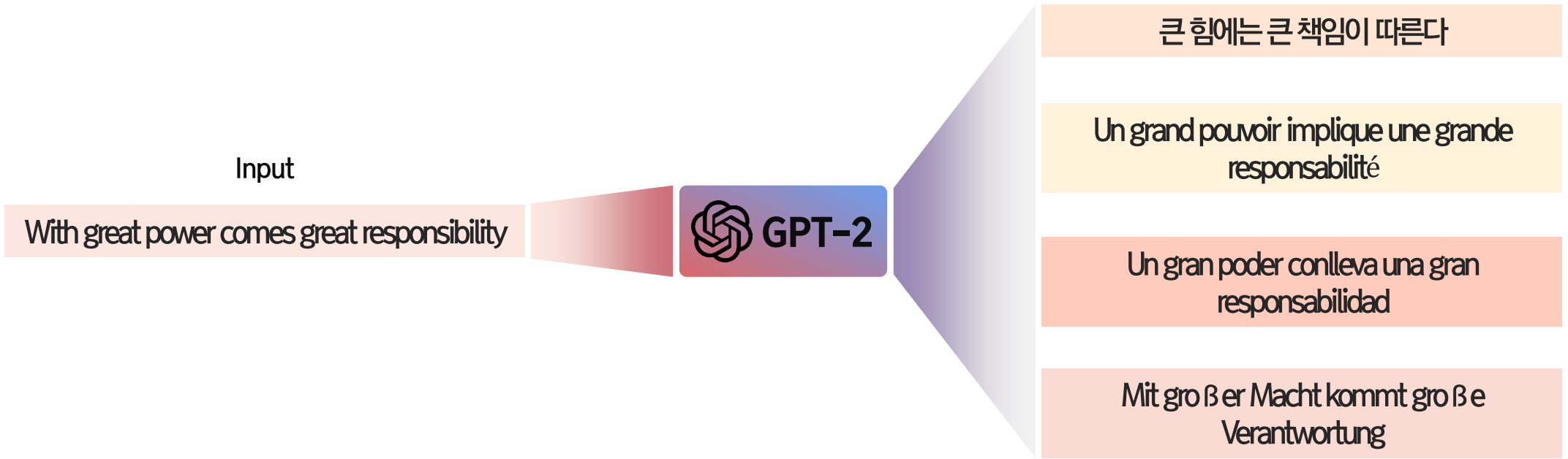
Large Language Model (LLM)

GPT Series : GPT-2

❖ Task Conditioning in GPT-2

- GPT-1 의 Objective : $P(Output|Input)$
- GPT-2 는 동일한 Unsupervised Model 을 사용하여 Multiple Tasks 학습을 목표
  - ✓ Objective : $P(Output|Input, Task)$

다른 Task 에 대해 동일한 Input 임에도 다른 Output 생성



Large Language Model (LLM)

GPT Series : GPT-2

❖ Zero-Shot Learning & Zero-Shot Task Transfer

- GPT-1 : Task의 종류에 따라 Input 을 Ordered Sequence 로 변환
- GPT-2 : 모델이 주어진 Task의 특성을 이해하고 알아서 답을 내릴 것으로 예상

With great power comes great responsibility

큰 힘에는 큰 책임이 따른다

With great power comes great responsibility

Un grand pouvoir implique une grande responsabilité

Peter and Elizabeth took a taxi to attend the night party in the city. While in the part, Elizabeth collapsed and was rushed to the hospital.

Elizabeth was hospitalized after attending a party with Peter.

Translate
English → Korean

Translate
English → French

Summarization

Large Language Model (LLM)

GPT Series : GPT-2

❖ Zero-Shot Learning & Zero-Shot Task Transfer

- GPT-1 : Task의 종류에 따라 Input 을 Ordered Sequence 로 변환
- GPT-2 : 모델이 주어진 Task의 특성을 이해하고 알아서 답을 내릴 것으로 예상

충분한 용량을 가진 언어모델은 더 나은 예측을 위해

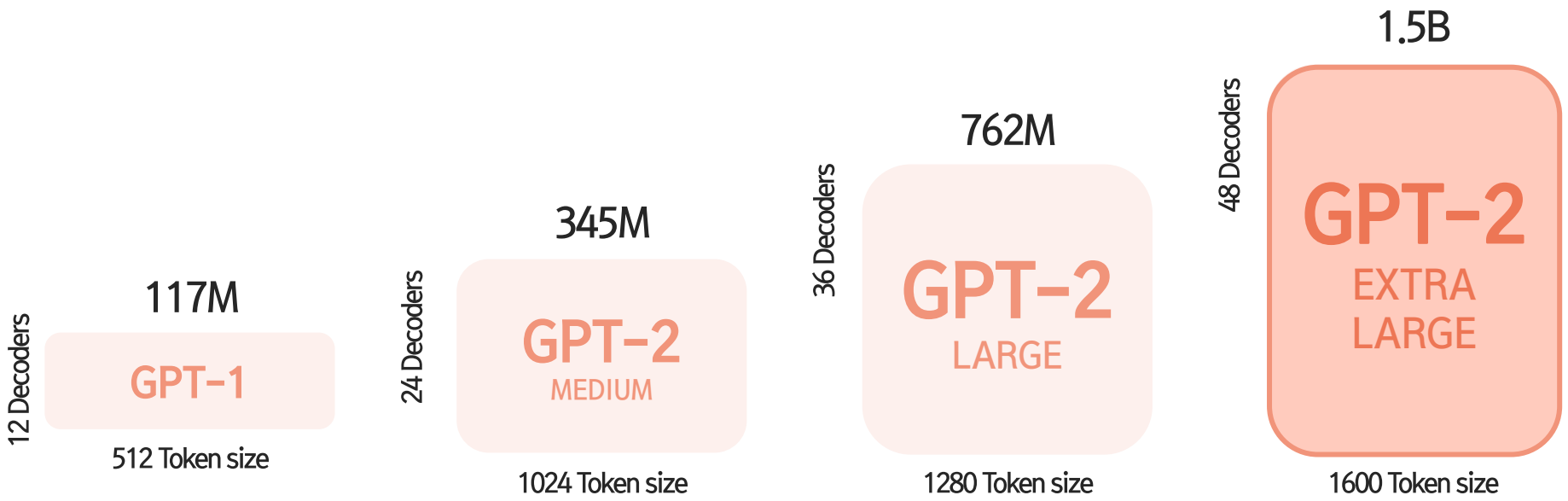
NL Sequences 에서 시연된 Task 를 유추하고 수행하기 시작할 것

Large Language Model (LLM)

GPT Series : GPT-2

❖ Language Models are Unsupervised Multitask Learners (2019)

- Parameters 의 개수가 가장 많은 모델 이 모든 Downstream Task 에서 가장 높은 성능
- 더 큰 Dataset & 더 많은 Parameters 학습을 통해 더 강력한 언어모델 학습 가능성 확인
  - ✓ 더 큰 Language Model 을 구축하자!

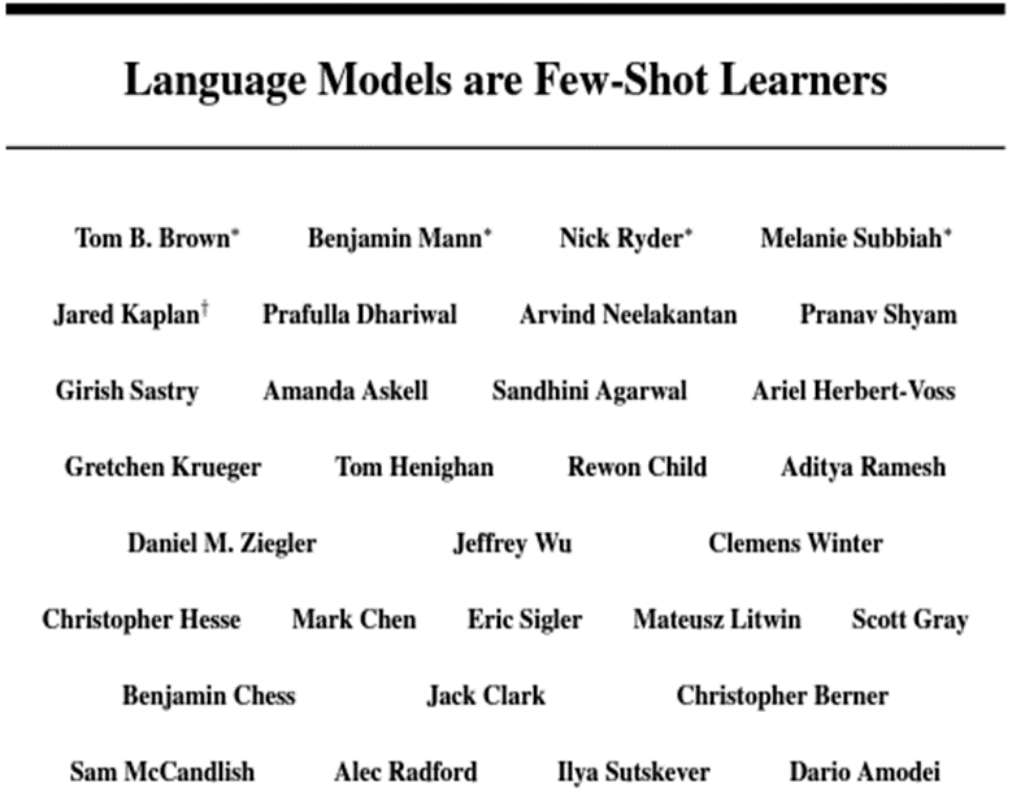


Large Language Model (LLM)

GPT Series : GPT-3

❖ Language Models are Few-Shot Learners (2020)

- 1,750 억 개의 파라미터를 가지며, 미세조정이 생략된 언어모델
- Few-shot & Zero-shot 방식으로 LLM 을 활용하는 In-context Learning 을 공식적으로 도입



Language Models are Few-Shot Learners

Tom B. Brown* Benjamin Mann* Nick Ryder* Melanie Subbiah*

Jared Kaplan† Prafulla Dhariwal Arvind Neelakantan Pranav Shyam

Girish Sastry Amanda Askell Sandhini Agarwal Ariel Herbert-Voss

Gretchen Krueger Tom Henighan Rewon Child Aditya Ramesh

Daniel M. Ziegler Jeffrey Wu Clemens Winter

Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray

Benjamin Chess Jack Clark Christopher Berner

Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei

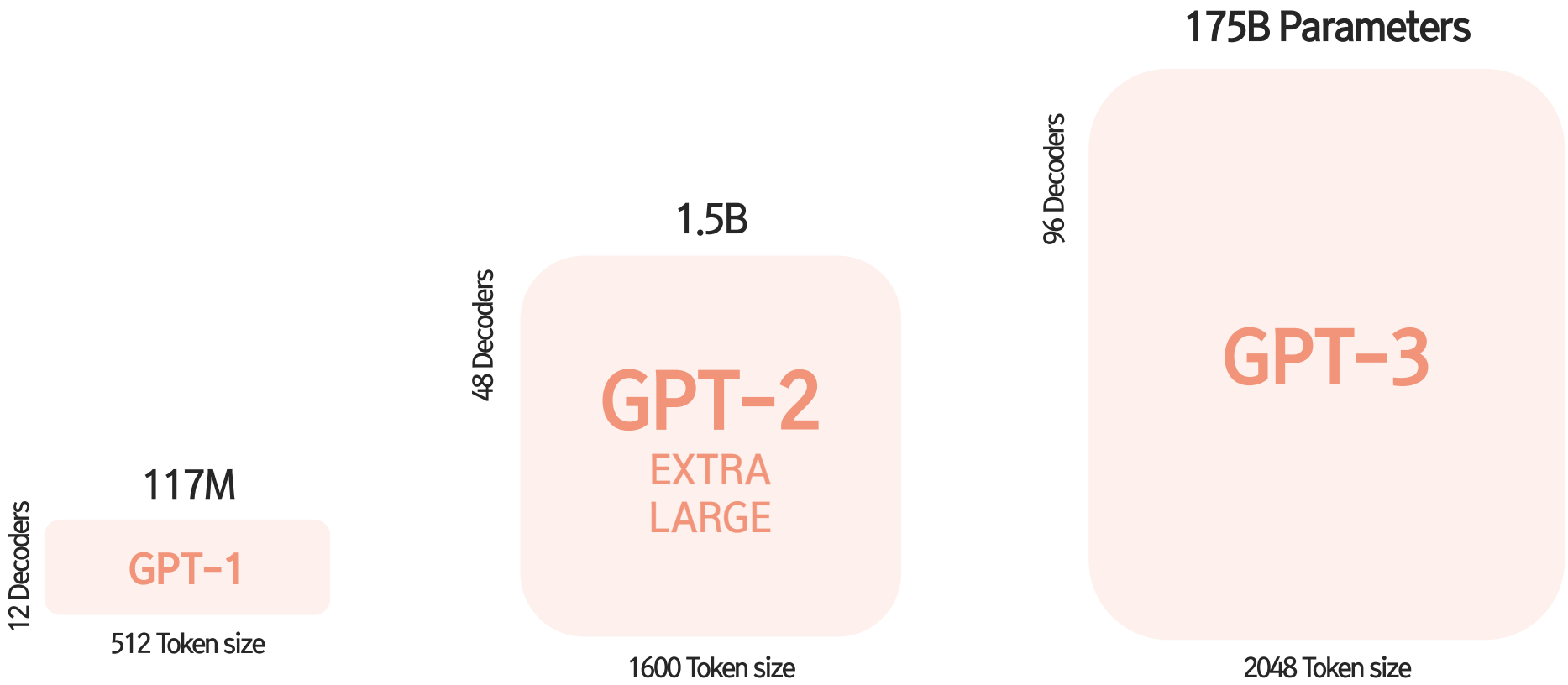
[8] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

Large Language Model (LLM)

GPT Series : GPT-3

❖ GPT-3 Architecture

- 더 많은 데이터, 더 깊은 구조로, 더 오랫동안 학습
  - ✓ GPT-2와 거의 유사한 구조를 가지며 우수한 성능 도출



Large Language Model (LLM)

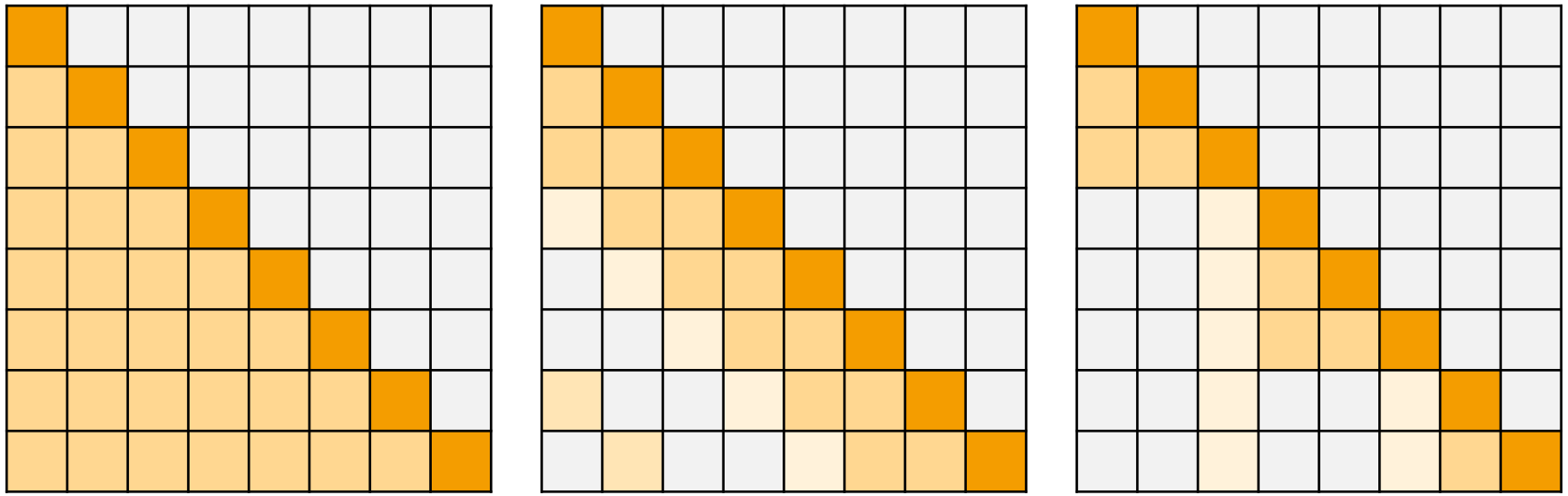
GPT Series : GPT-3

❖ GPT-3 Architecture

- 더 많은 데이터, 더 깊은 구조로, 더 오랫동안 학습
- Transformer Layer 에서 Dense & Locally Banded Sparse Attention 사용 (Sparse Transformer)
  - ✓ Locally banded : Noise와 같이 관련 없는 정보의 영향 ↓
  - ✓ Sparse : Attention Score 계산 횟수 ↓

Input Sequence 의 Long-term Dependencies 포착

모델의 효율성과 확장성 유지



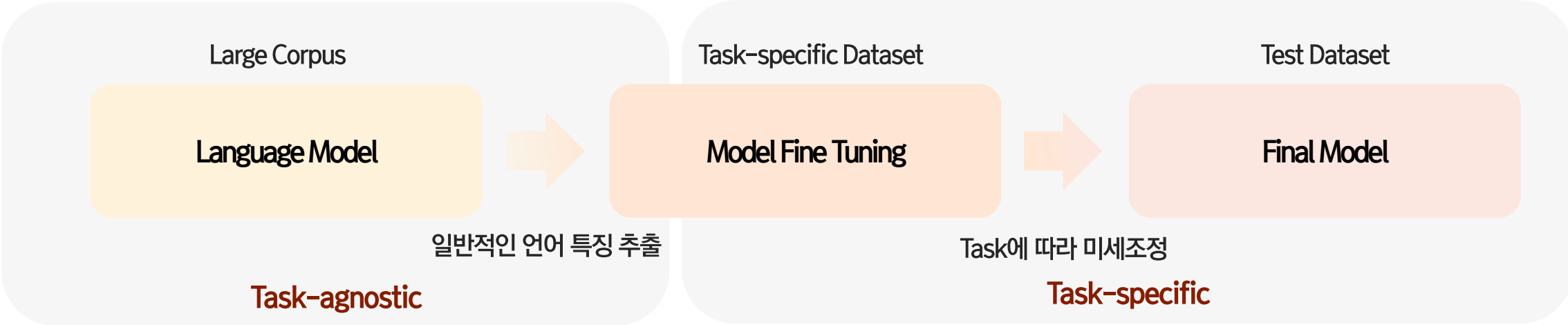
Transformer | Sparse Transformer (strided) | Sparse Transformer (fixed)

Large Language Model (LLM)

GPT Series : GPT-3

❖ Overall Process of Language Model

- 기존 사전 학습된 언어모델의 구조는 Task-agnostic 임에도 불구하고 미세조정 단계에서는 여전히 해당 Task의 Labeled Dataset 필요
- Task-specific 한 Dataset 을 바탕으로 Task-specific 미세조정 수행

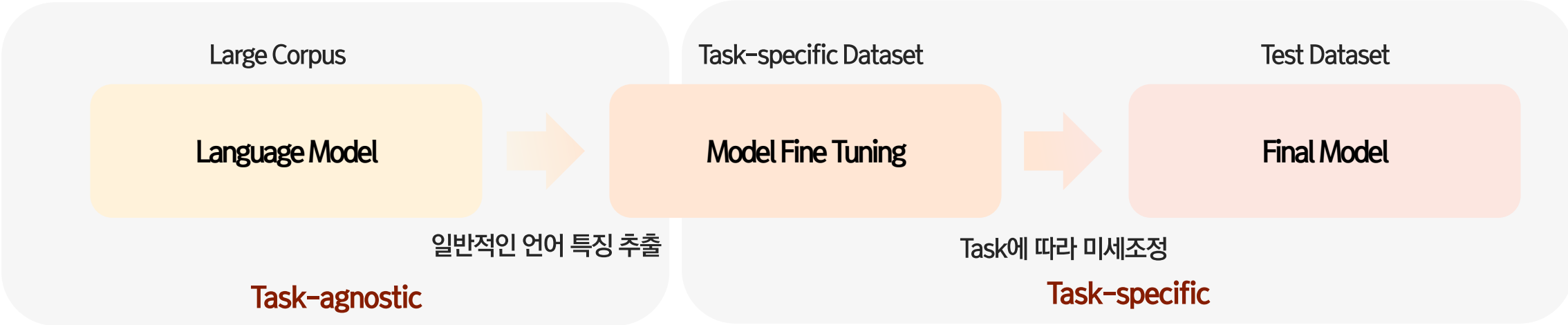


Large Language Model (LLM)

GPT Series : GPT-3

❖ Limitations of Fine Tuning

- 새로운 Task에 대한 Task 별 Labeled Dataset 필요
- 미세조정을 거쳐 작은 Task의 분포를 따르면 과적합이 발생하여 분포 밖 Dataset 에 대해 일반화 성능을 잃을 수 있음
- 사람과 같이 적은 샘플로도 Task를 수행하는 능력 필요

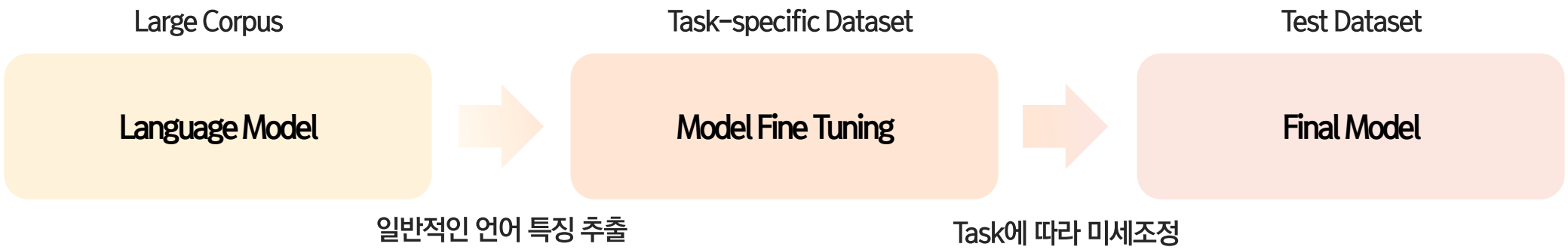


Large Language Model (LLM)

GPT Series : GPT-3

❖ GPT-3 : No Fine-tuning Required

- Labeled Dataset 이 필요한 미세조정 단계를 제외
- 미세조정 없이 사전학습 만으로 동작하는 모델 구축

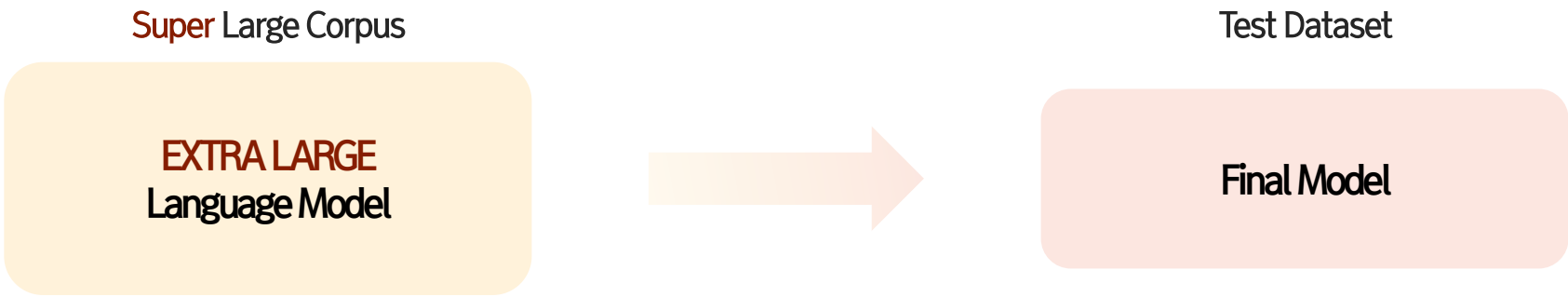


Large Language Model (LLM)

GPT Series : GPT-3

❖ GPT-3 : No Fine-tuning Required

- Labeled Dataset 이 필요한 미세조정 단계를 제외
- 미세조정 없이 사전학습 만으로 동작하는 모델 구축
  - ✓ In-context Learning
  - ✓ Large Capacity

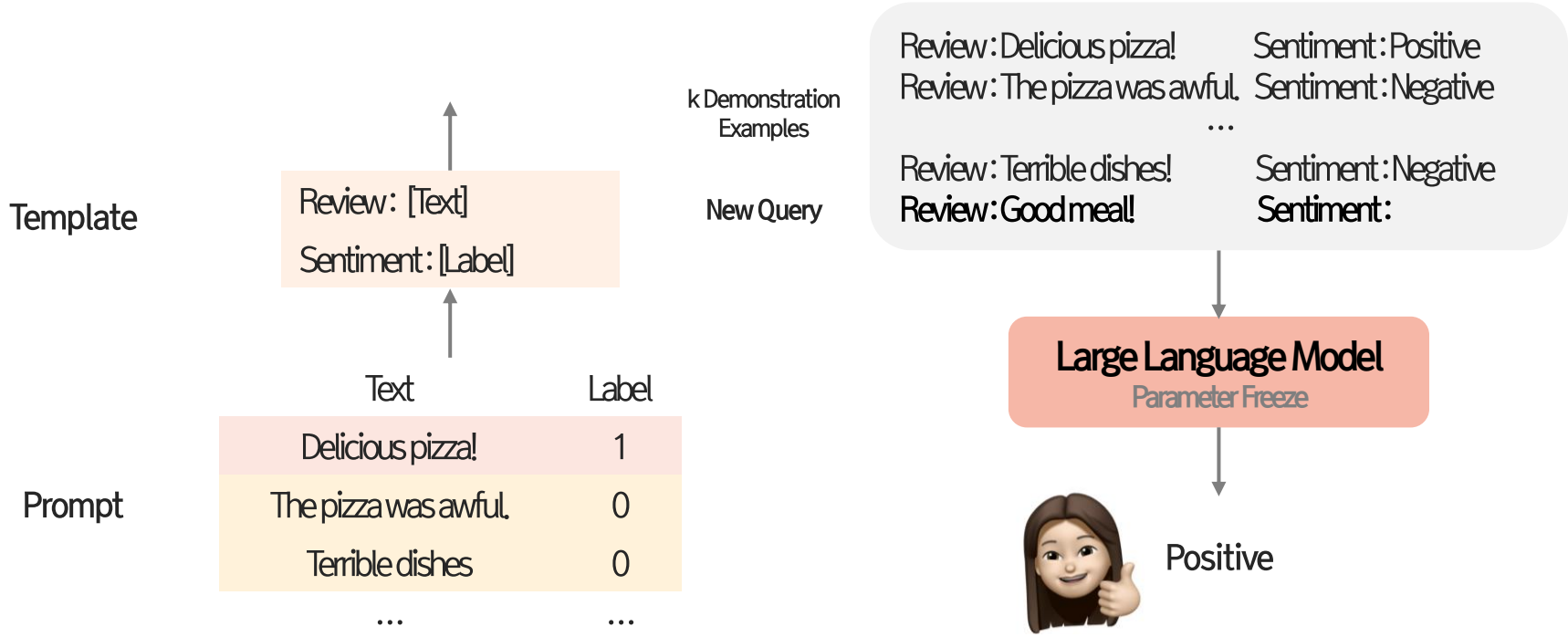


Large Language Model (LLM)

GPT Series : GPT-3

❖ In-context Learning

- 프롬프트의 맥락적 의미 (In-context)를 모델이 이해하고, 이에 대한 답변 생성
- 사전학습이나 미세조정처럼 모델을 업데이트하지 않고, 별도의 모델 학습 과정 없음
- Demo Context 구성을 위한 소수의 예시 필요



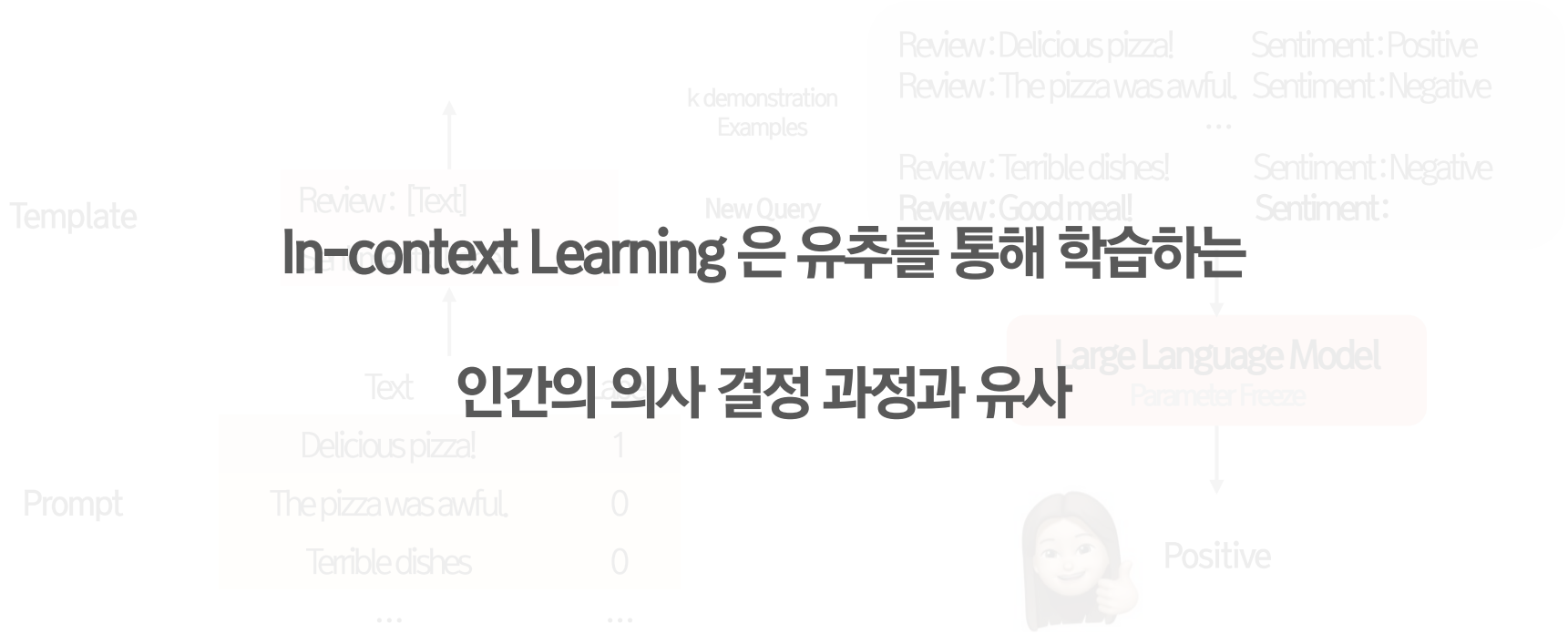
https://medium.com/@mlblogging.k/understanding-in-context-learning-in-large-language-models-like-gpt3-gpt-j-gptneox-e0a71063a6db

Large Language Model (LLM)

GPT Series : GPT-3

❖ In-context Learning

- 프롬프트의 맥락적 의미 (In-context)를 모델이 이해하고, 이에 대한 답변 생성
- 사전학습이나 미세조정처럼 모델을 업데이트하지 않고, 별도의 모델 학습 과정 없음
- Demo Context 구성을 위한 소수의 예시 필요



In-context Learning 은 유추를 통해 학습하는

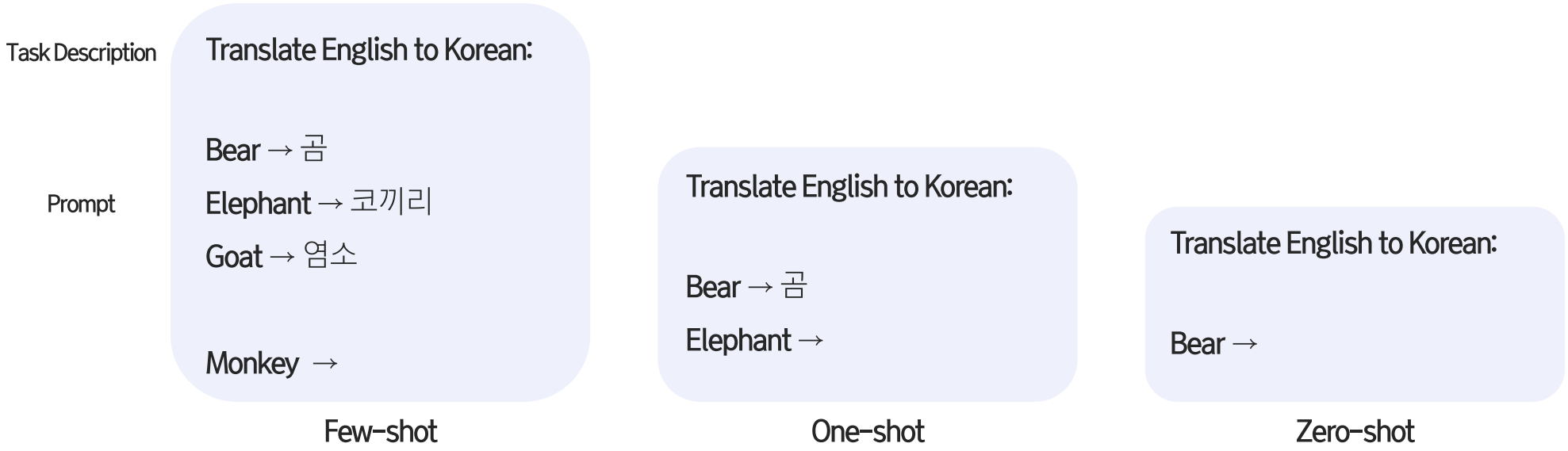
인간의 의사 결정 과정과 유사

Large Language Model (LLM)

GPT Series : GPT-3

❖ Few-shot, One-shot and Zero-shot

- Few-shot : 모델에게 Task 설명 + Model 의 Context Window k에 맞는 예제 제공 (Limited Labeled Dataset)
- One-shot : 모델에게 하나의 예제만 제공 (One Labeled Example)
- Zero-shot : 모델에게 예시 제공 없음, 사전 지식만으로 예측 (No Labeled Example)

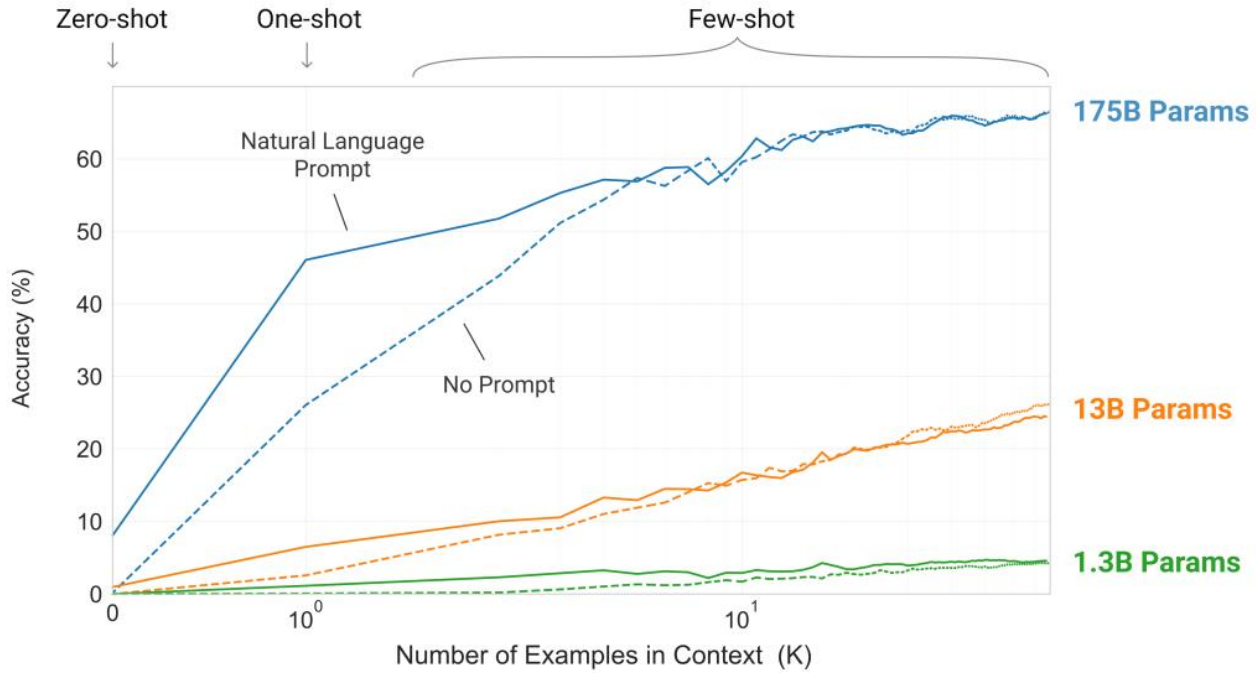


Large Language Model (LLM)

GPT Series : GPT-3

❖ Few-shot, One-shot and Zero-shot

- Context Window k 가 증가할 수록 성능 향상 확인
- 모델이 커질수록 In-context Learning 의 효율 증가
- Task 에 대한 Prompt 의 도움 확인



[8] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

Large Language Model (LLM)

GPT Series : GPT-3

❖ Limitations of GPT-3

- 긴 문장 구성에 일관성을 잃고 Text Sequence 반복하는 경향
- 모델의 단방향성으로 인해 자연어 추론과 같은 Task 를 잘 해내지 못함
  - ✓ 유사한 규모의 양방향 모델 고려 가능
- 각 Token 에 동일한 가중치가 부여되므로 목표 지향적 예측 개념 결여

Large Language Model (LLM)

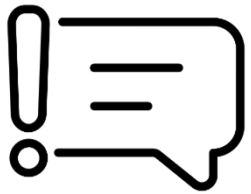
GPT Series : GPT-3

❖ Broader Impacts of GPT-3

- 언어모델의 오용 : 가짜뉴스, 스팸, 피싱 등
- 공정성, 편향성 및 대표성 : 인터넷을 통해 수집한 데이터를 기반으로 하기 때문에 그곳에 존재하는 편견과 편향을 그대로 학습
  - ✓ 성별, 민족, 인종, 종교 에 대한 편견과 선입견 존재
- 에너지 소비 : 모델 학습에는 상당한 자원을 소비하지만, 훈련 후에는 효율적
  - ✓ 모델을 한번 훈련시키는데 약 50억 소요, 하나의 Tesla V100 사용시 355년 소요

미래의 영향을 예측하기 어려움

불쾌한 컨텐츠 생성



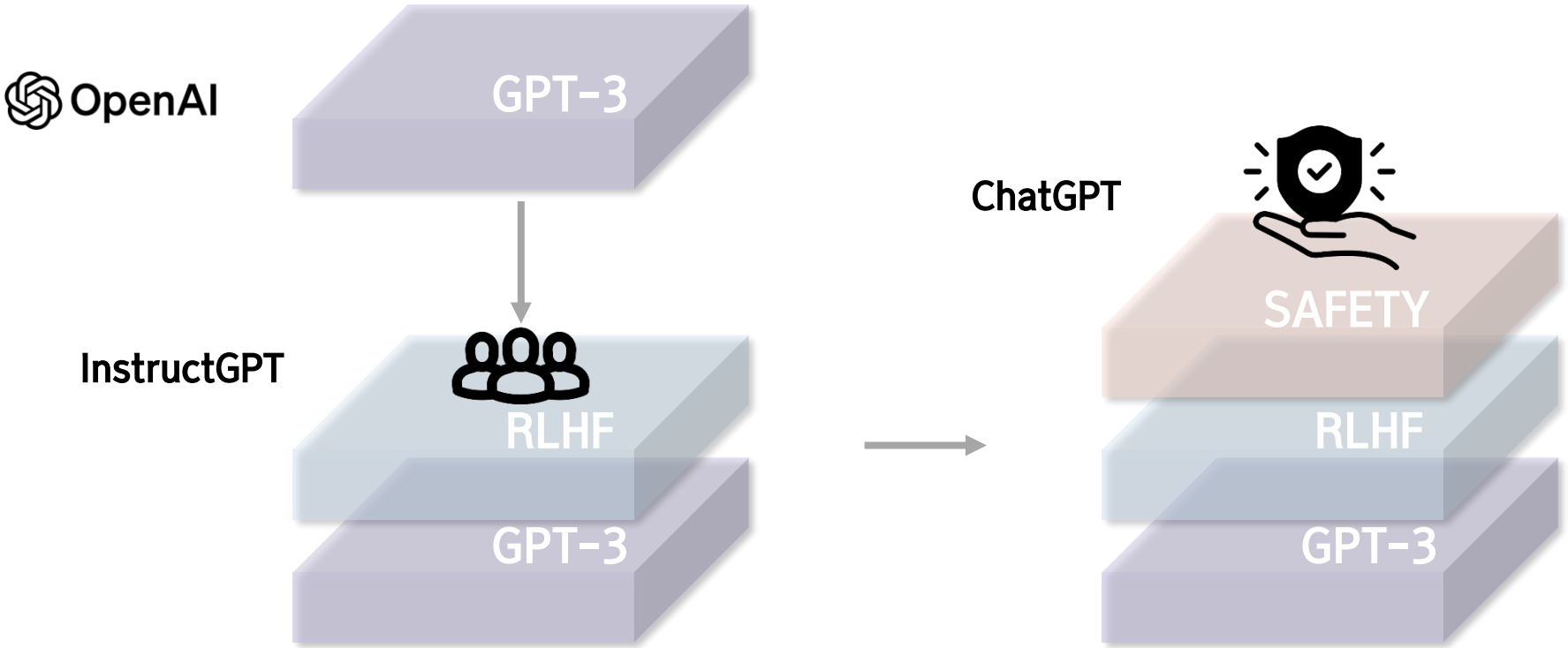
아직 모든 것을 잘하진 않음

4. ChatGPT

ChatGPT

❖ ChatGPT

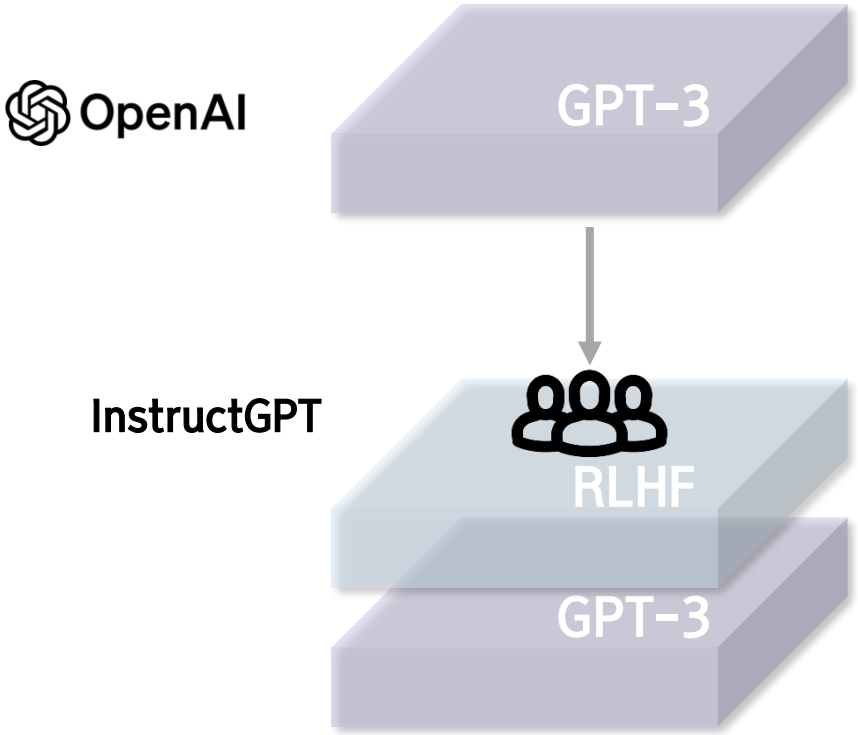
- 2022년 11월에 OpenAI 에서 출시한 대화형 인공지능 챗봇
  - ✓ GPT-3가 개선된 GPT-3.5 및 GPT-4 모델 기반
- Instruct GPT 와 유사한 방식으로 학습되었지만, 대화에 최적화 & 안전 강화

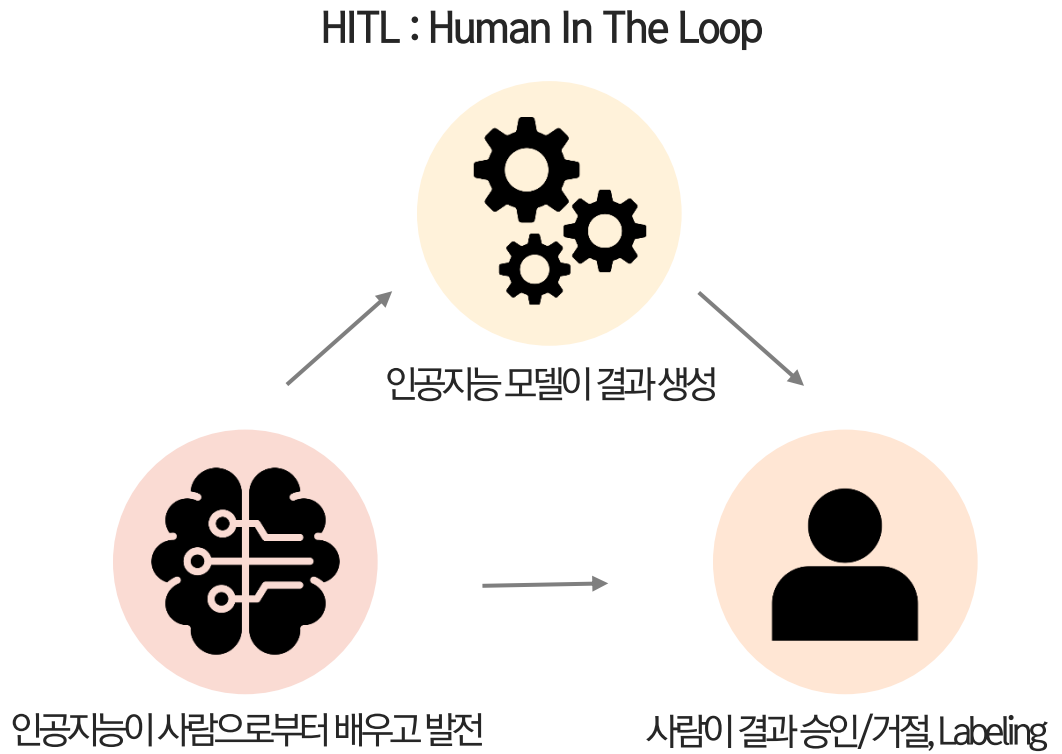


ChatGPT

❖ Instruct GPT (2022.03)

- GPT-3의 방향이 사용자의 의도와 부합하지 않는 단점 : Untruthful, Toxic, Not Helpful
- 인간 피드백 기반 강화학습을 기존 GPT-3 에 도입하여 사용자의 지시를 잘 수행하도록 함

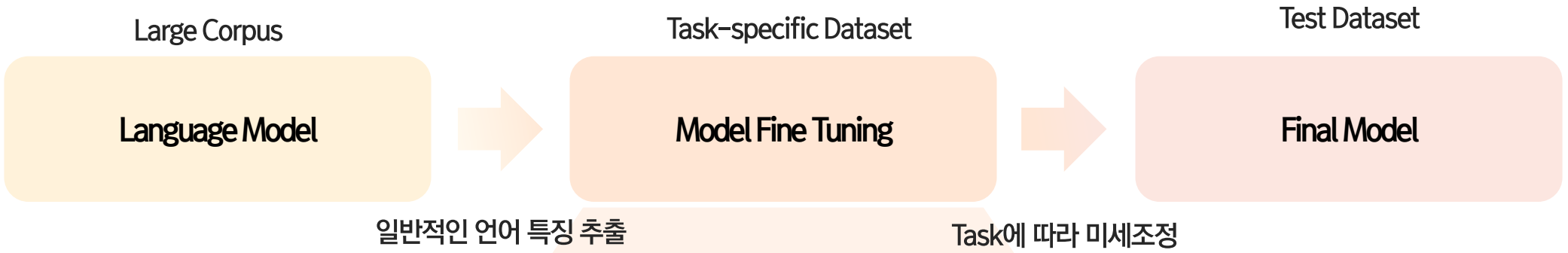




ChatGPT

❖ RLHF : Reinforcement Learning from Human Feedback

- 인간이 작성한 설명과 인간이 판단한 Output 의 Ranking 을 모델에 반영
- 인간의 선호도를 보상 (Reward Signal) 으로 사용하여 미세조정



40명의 Crowd worker 고용

Step 1 : Supervised Policy 학습

Step 2 : Reward Model 학습

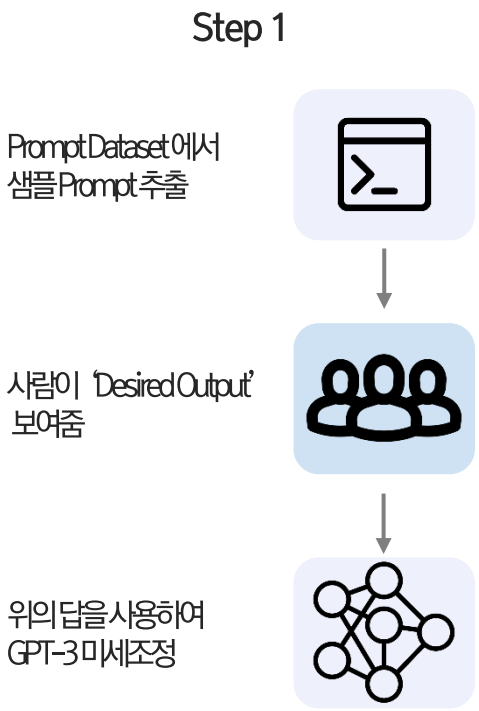
Step 3 : 강화학습을 통한 Policy 최적화

사용자의 지시 잘 수행 & Toxic Output 감소

ChatGPT

❖ Instruct GPT (2022.03)

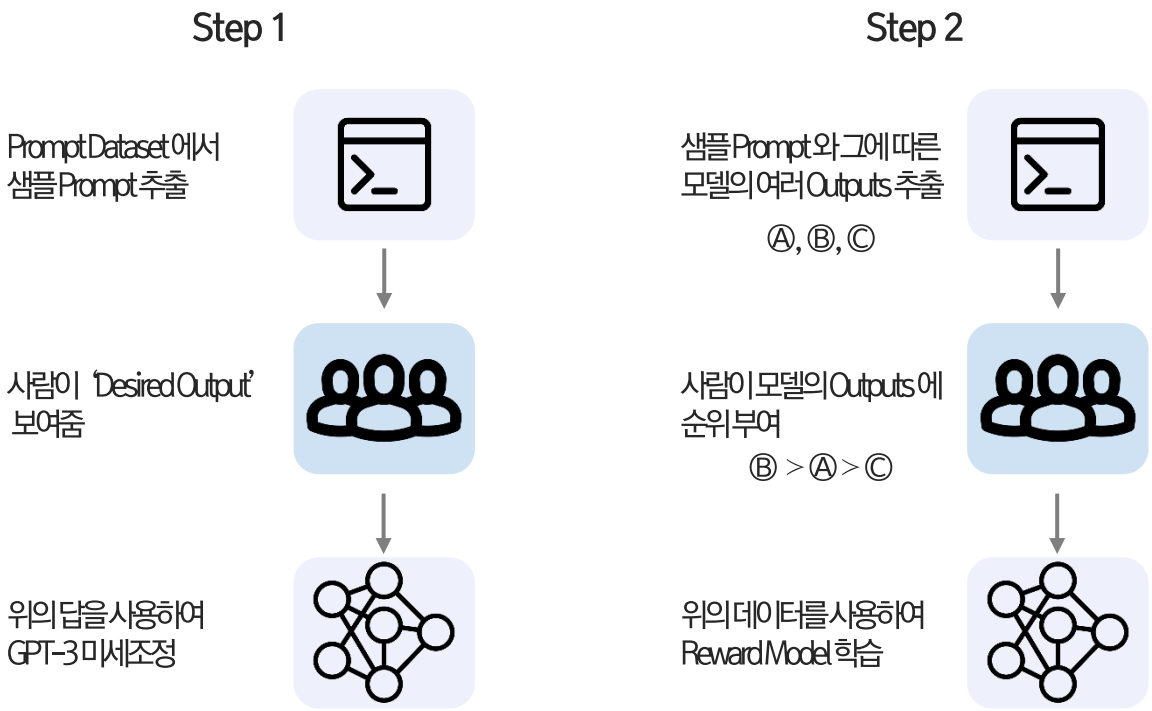
- Step 1 : Supervised Policy 학습
  - ✓ 사람이 GPT 인척 하며 'Desired Output' 데이터 생성
  - ✓ 생성된 데이터를 사용해서 GPT-3의 미세조정 진행



ChatGPT

❖ Instruct GPT (2022.03)

- Step 2 : Reward Model 학습
  - ✓ 사람이 GPT의 Output 에 순위 부여
  - ✓ Ranked Data를 통해 사람이 어떤 Output 을 더 선호할지 예측하는 Reward Model 학습



ChatGPT

❖ Instruct GPT (2022.03)

- Step 3 : 강화학습을 통한 Policy 최적화
  - ✓ 강화학습 : Agent 가 Environment 와 상호작용 하며 시행착오를 통해 Reward를 최대화하는 방법을 학습

Agent

Action A_t

State S_t

Reward R_{t+1}

Environment

ChatGPT

❖ Instruct GPT (2022.03)

- Step 3 : 강화학습을 통한 Policy 최적화
  - ✓ 강화학습 : Agent 가 Environment 와 상호작용 하며 시행착오를 통해 Reward를 최대화하는 방법을 학습

Agent

Agent : 학습하고자 하는 모델

Environment : 주변 환경

Action : 모델이 취하는 행동

Policy : 모델의 행동을 결정하는 함수

Reward : 모델의 행동에 대해 환경이 주는 반응

Environment

ChatGPT

❖ Instruct GPT (2022.03)

- Step 3 : 강화학습을 통한 Policy 최적화
  - ✓ 강화학습 : Agent 가 Environment 와 상호작용 하며 시행착오를 통해 Reward를 최대화하는 방법을 학습

Agent

Agent : GPT

Environment : Input

Action : Generated Output

Policy : Parameters of GPT

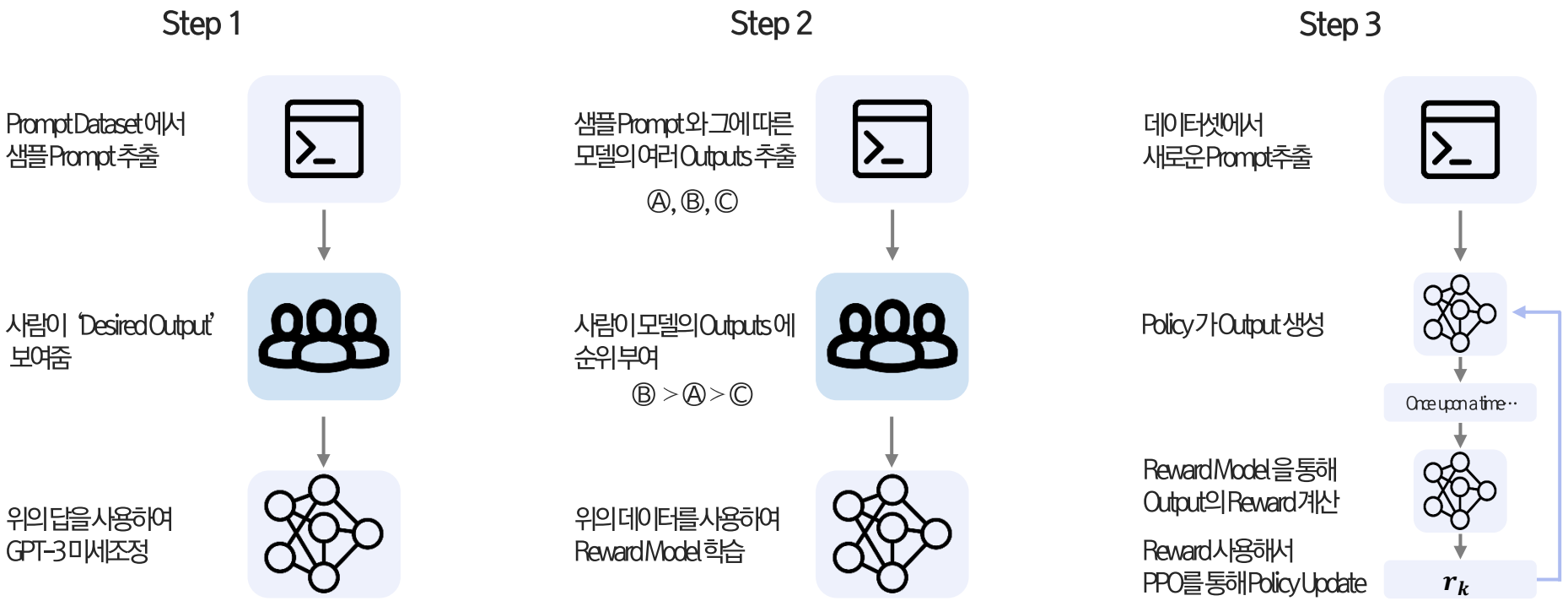
Reward : Step 2에서 사람이 부여한 Reward

Environment

ChatGPT

❖ Instruct GPT (2022.03)

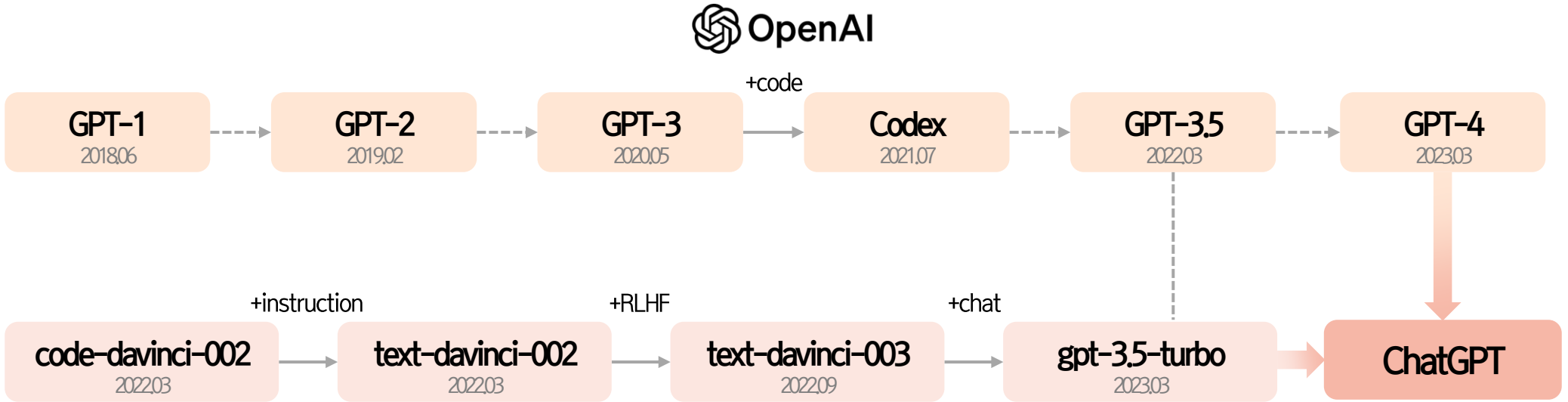
- Step 3 : 강화학습을 통한 Policy 최적화
  - ✓ 앞서 학습한 Reward Function 을 최대화 하기 위해 GPT-3의 Policy 미세조정 진행
  - ✓ PPO를 통해 GPT-3의 Policy 학습



ChatGPT

❖ Technical Evolution of GPT Series Models

- 블로그 기사 및 OpenAI 의 공식 API 를 기반으로 나타낸 GPT Series 모델의 기술적 진화
  - ✓ 실선 : 두 모델 간 진화경로에 대한 명시적 증거, 점선 : 상대적으로 약한 진화 관계

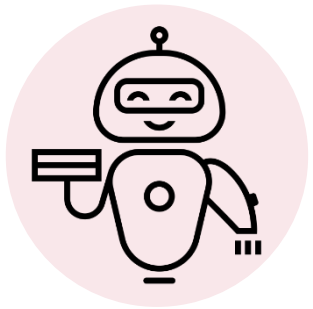


[1] Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223.

ChatGPT

❖ After ChatGPT

- 텍스트를 행동으로 전환하는 AI
- 코드를 작성하는 AI
- GPT Upgrade 와 인공지능 경쟁



Text-Action Transformer

Writing Code

AI Arm's Race

https://medium.com/@sgunnisonmiller/whats-next-after-chatgpt-four-looming-advancements-926aa0d51bcb

5. Conclusion

Conclusions

What is LLM and ChatGPT?

❖ From Transformer to Large Language Model

- Transformer 의 Encoder 기반 – BERT 계열 / Decoder 기반 – GPT 계열

❖ Large Language Model (LLM)

- 방대한 텍스트 데이터로 학습된 수천 억 개의 파라미터를 가진 모델로 언어를 이해하고 복잡한 과업을 수행하는데 강력한 능력

❖ GPT Series

- GPT-1 : Transformer 의 Decoder 를 기반으로 비지도 사전 훈련과 지도 미세조정을 결합한 생성 방식의 언어 모델 학습
- GPT-2 : 명시적인 Task 구분 없이 모델이 주어진 Input 통해 적합한 Task 를 유추
- GPT-3 : 1,750 억 개의 파라미터를 가지며 미세조정이 생략된 언어모델

❖ ChatGPT

- 기존 GPT-3 에 인간 피드백 기반 강화학습을 도입한 대화형 인공지능 챗봇

6. Reference

References

[1] Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223.
[2] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. Advances in neural information processing systems, 27.
[3] Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.
[4] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.
[5] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.
[6] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.
[7] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.
[8] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

Thank You