IISR 2023 Freshman Orientation

Day 4 : Pre-trained Model

薛竣祐 Alvin

智慧型資訊服務研發實驗室
Intelligent Information Service Research Lab

Outline

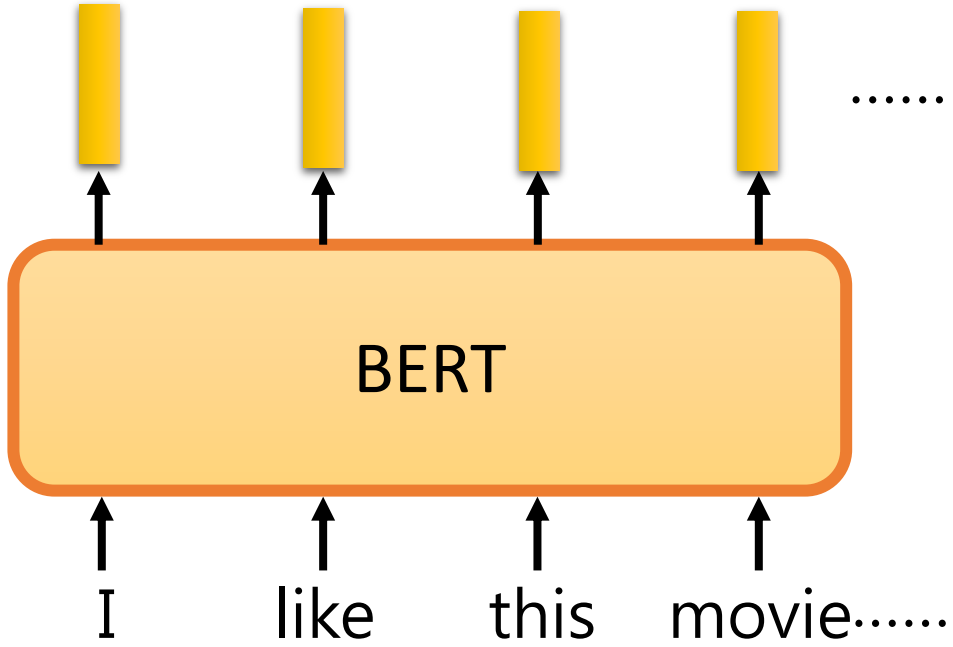
- What is Pre-trained Model & How to Pre-train
  (Use Bert as Example)
- Introduce various Pre-trained Models
  - GPT
  - ELECTRA
  - DeBERTa
  - Bart

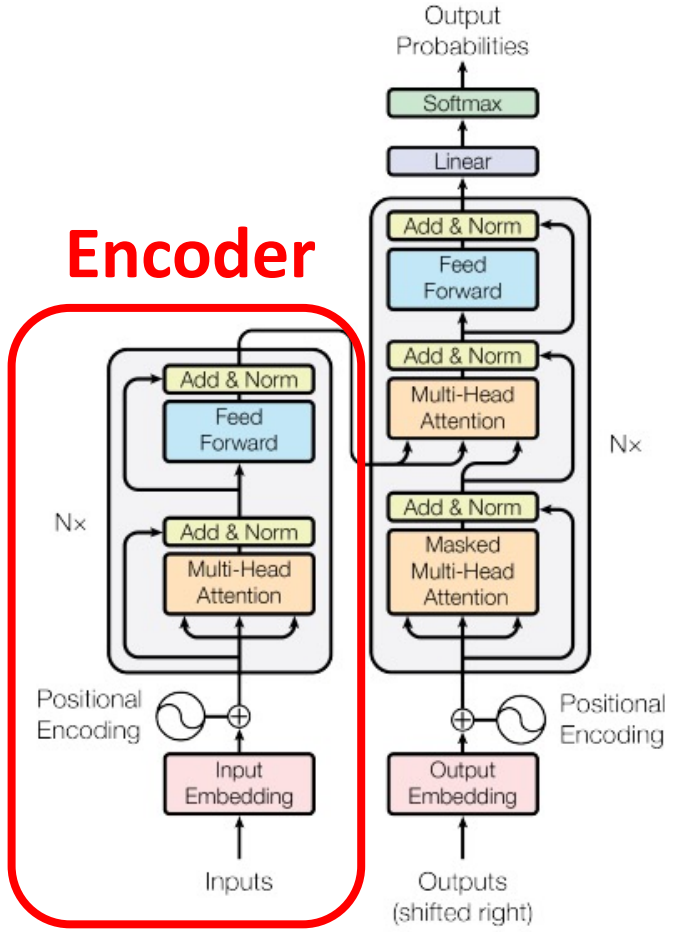
Modify From Hung-Yi Lee Machine Learning (2021)

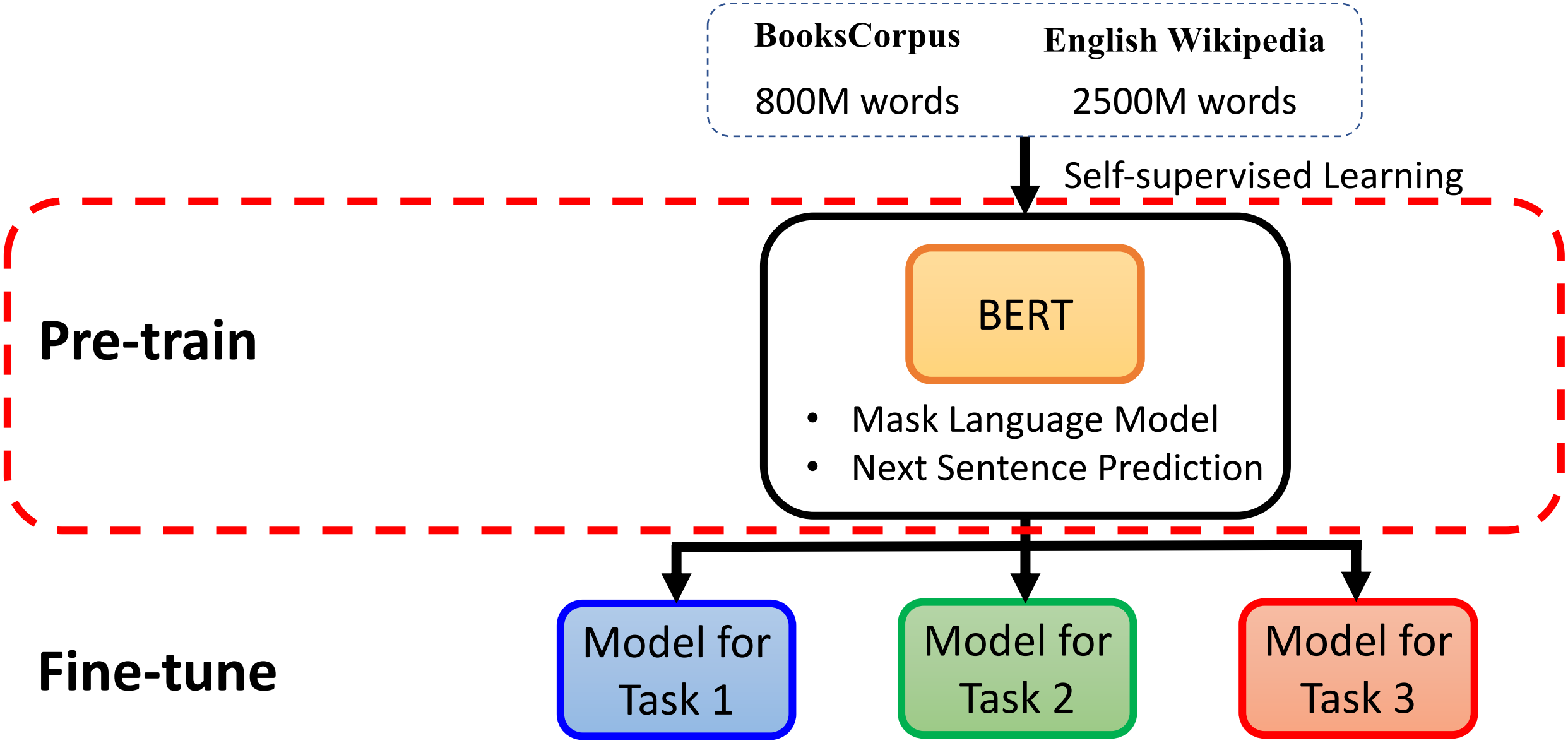
Pre-trained Model

Bidirectional Encoder Representations from Transformers (BERT)

BERT = **Encoder** of Transformer



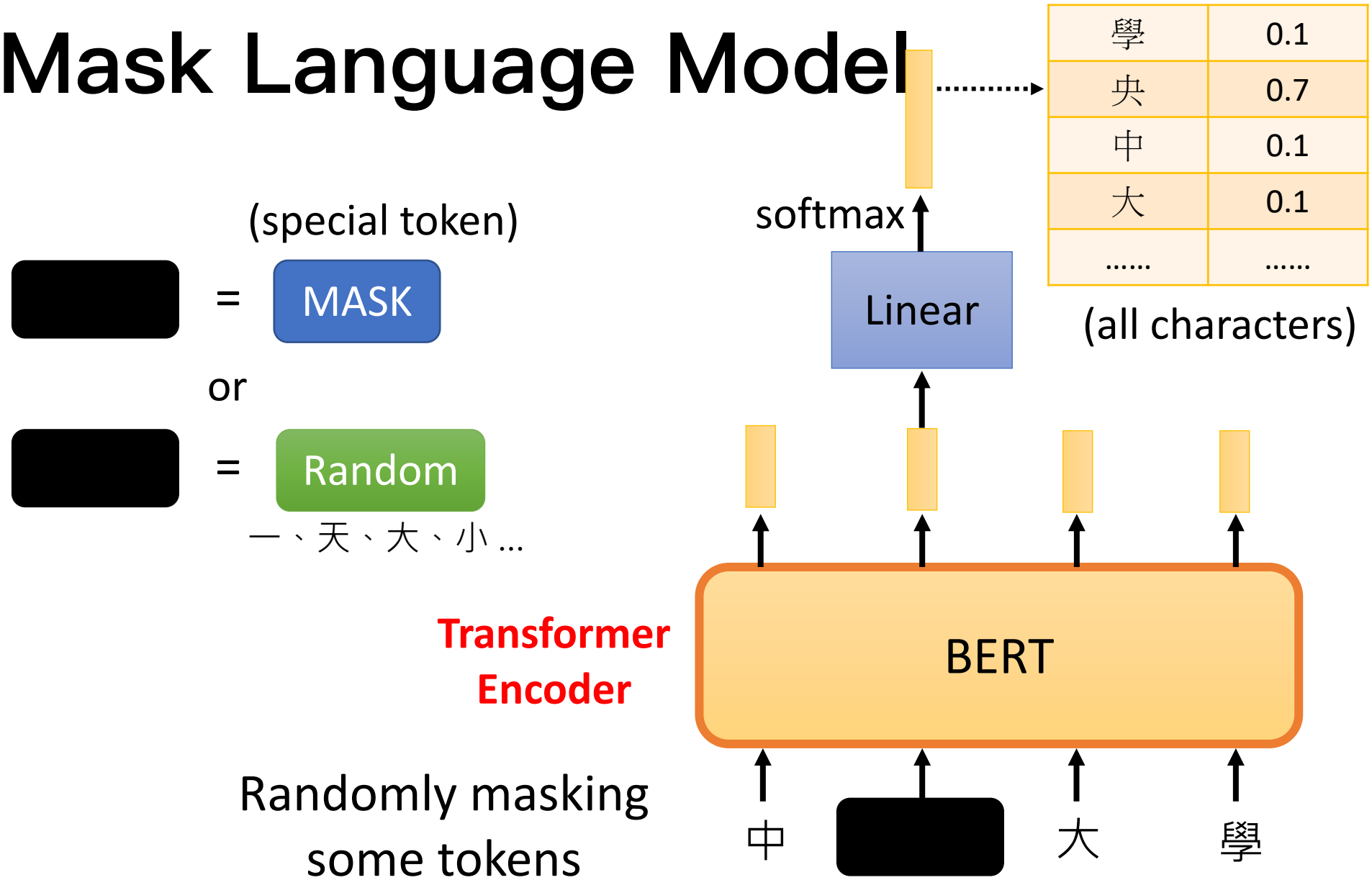




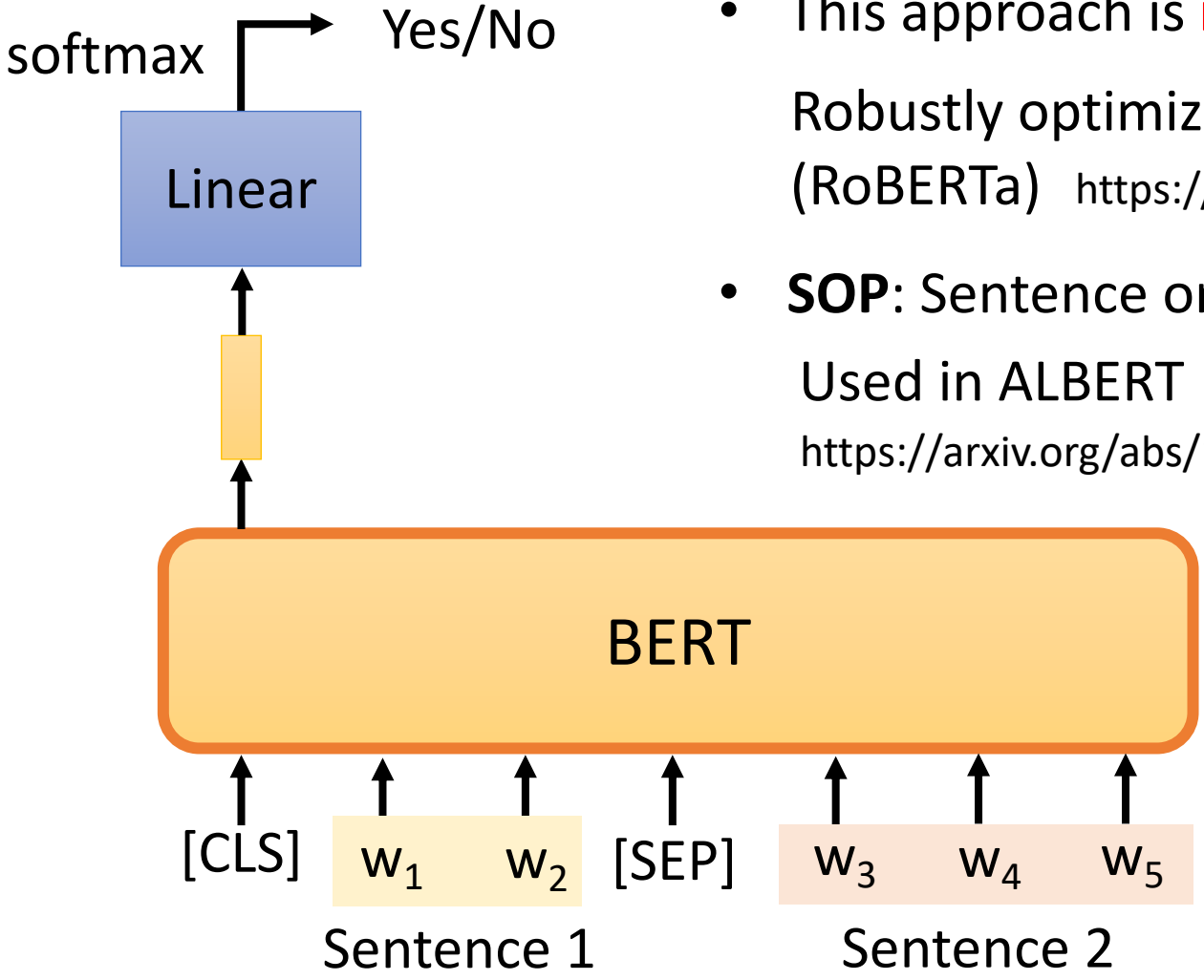
Downstream Tasks

- The tasks we care.
- We have a little bit labeled data.

Mask Language Model



Next Sentence Prediction



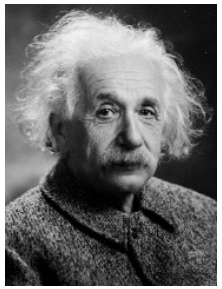
- This approach is **not helpful**.

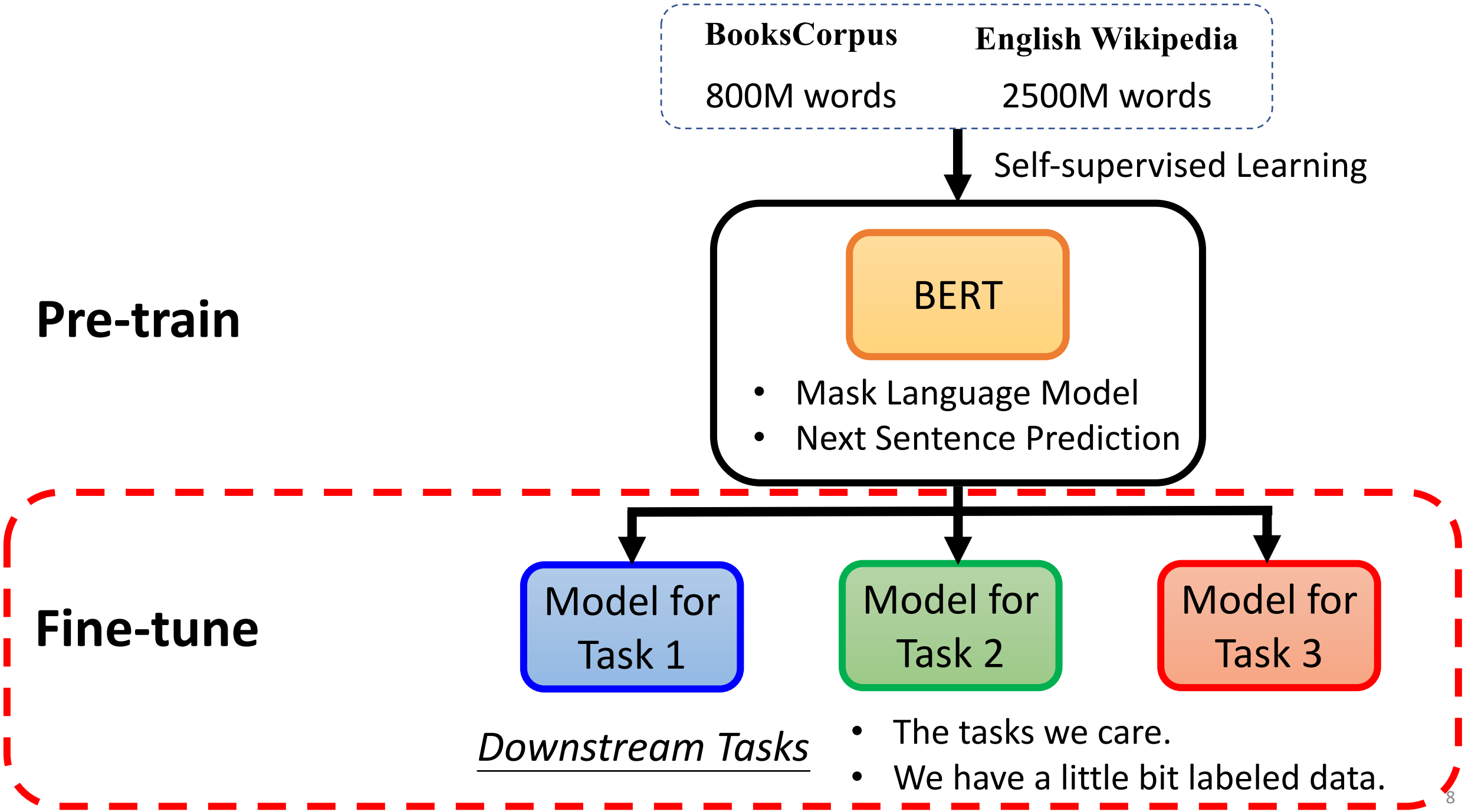
  Robustly optimized BERT approach (RoBERTa) https://arxiv.org/abs/1907.11692

- **SOP**: Sentence order prediction

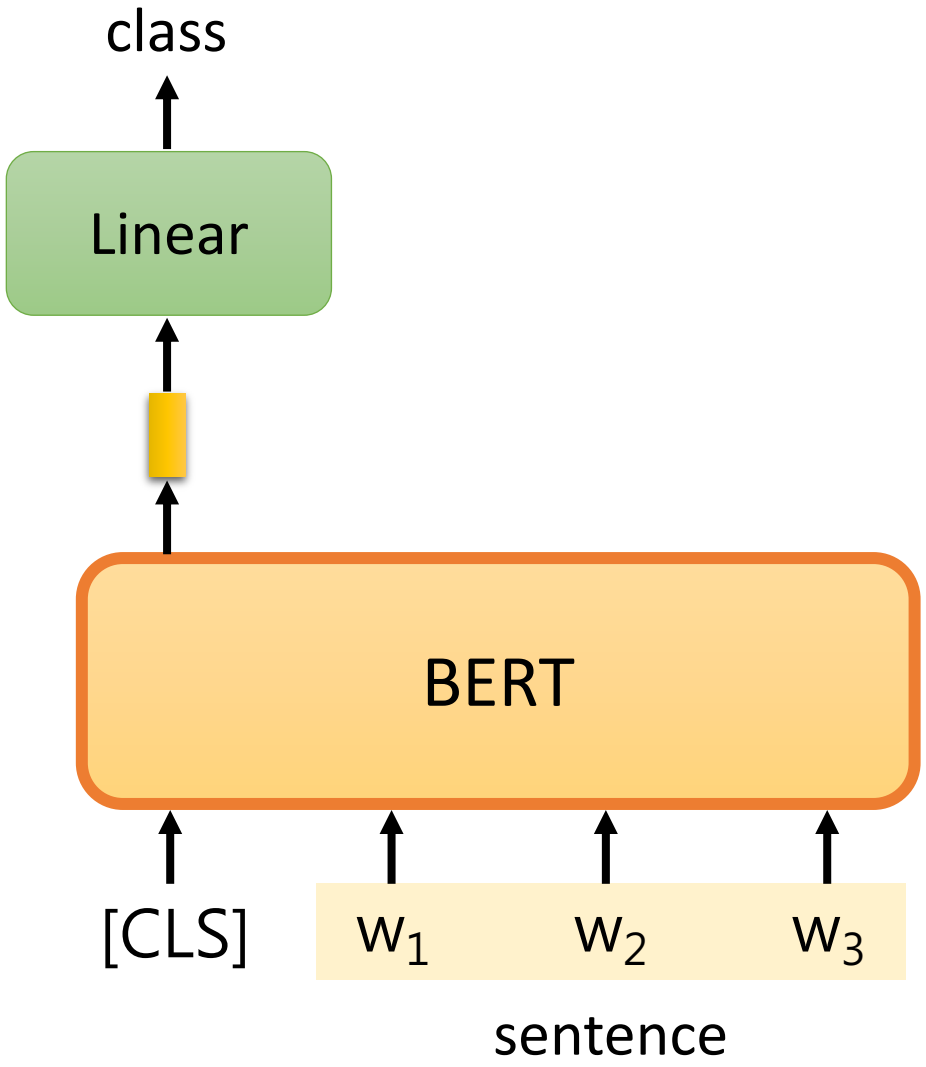
  Used in ALBERT

  https://arxiv.org/abs/1909.11942



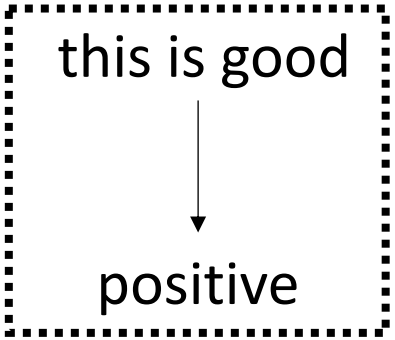

BooksCorpus
800M words
English Wikipedia
2500M words
Self-supervised Learning
Pre-train
BERT
• Mask Language Model
• Next Sentence Prediction
Fine-tune
Model for Task 1
Model for Task 2
Model for Task 3
Downstream Tasks
• The tasks we care.
• We have a little bit labeled data.

How to use BERT – Case 1

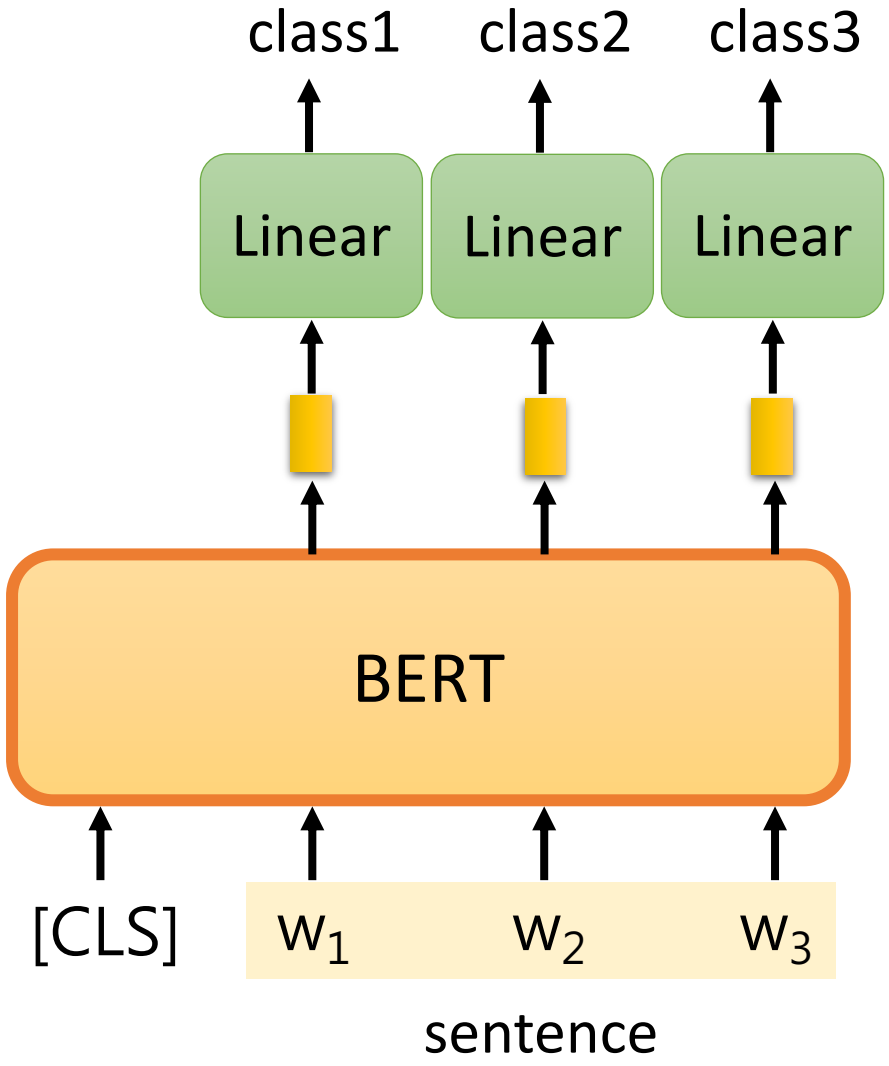


Input: sequence
output: class

Example:
Sentiment analysis

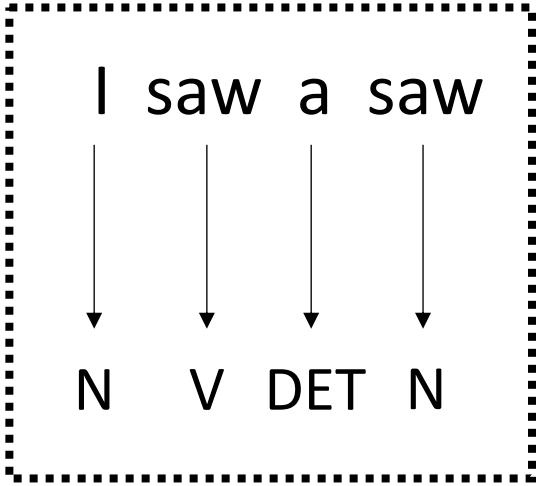


How to use BERT – Case 2



Input: sequence
output: same as input

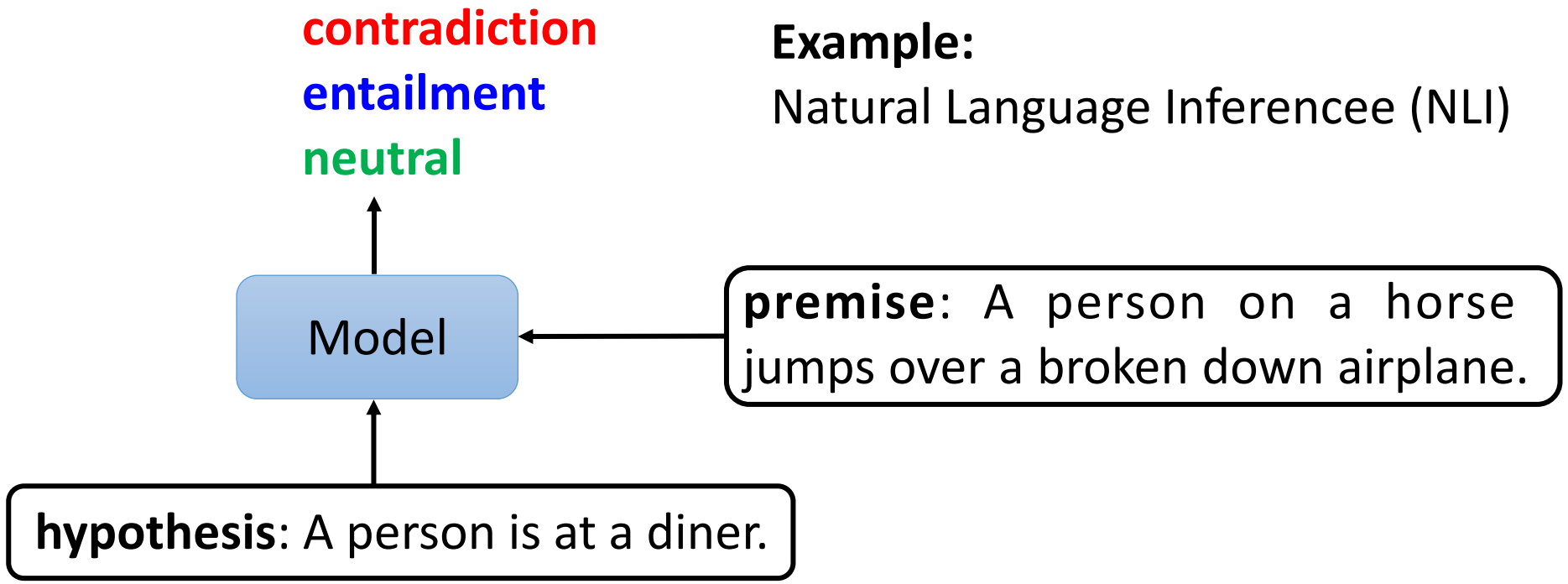
Example:
POS tagging



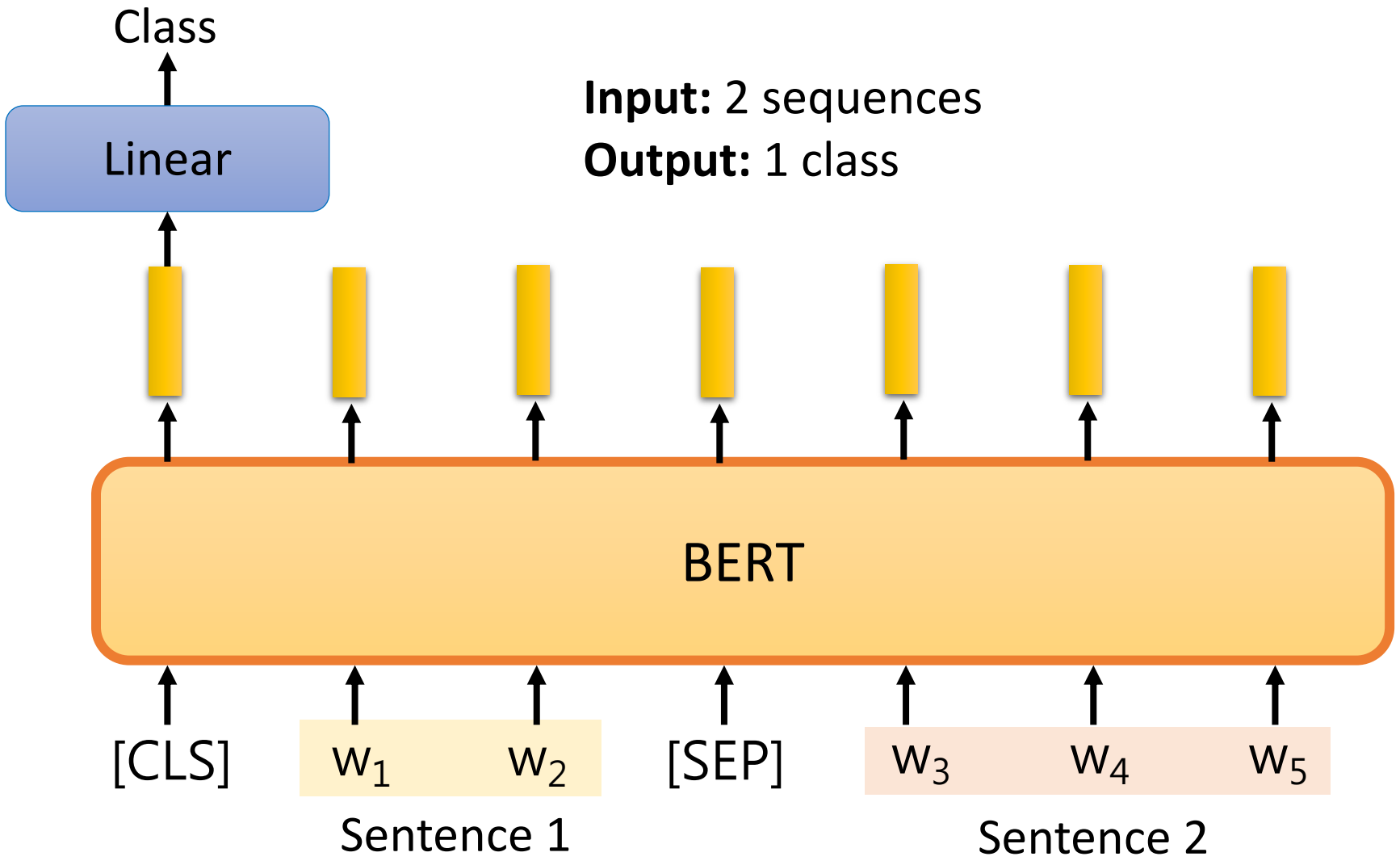
How to use BERT – Case 3

Input: 2 sequences
Output: 1 class

Example:
Natural Language Inferencee (NLI)



How to use BERT – Case 3

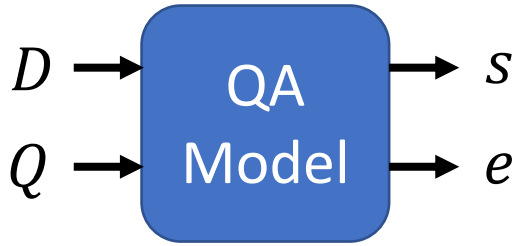


How to use BERT – Case 4

- Extraction-based Question Answering (QA)

Document: $D = \{d_1, d_2, \cdots, d_N\}$

Query: $Q = \{q_1, q_2, \cdots, q_M\}$



Answer: $A = \{d_s, \cdots, d_e\}$

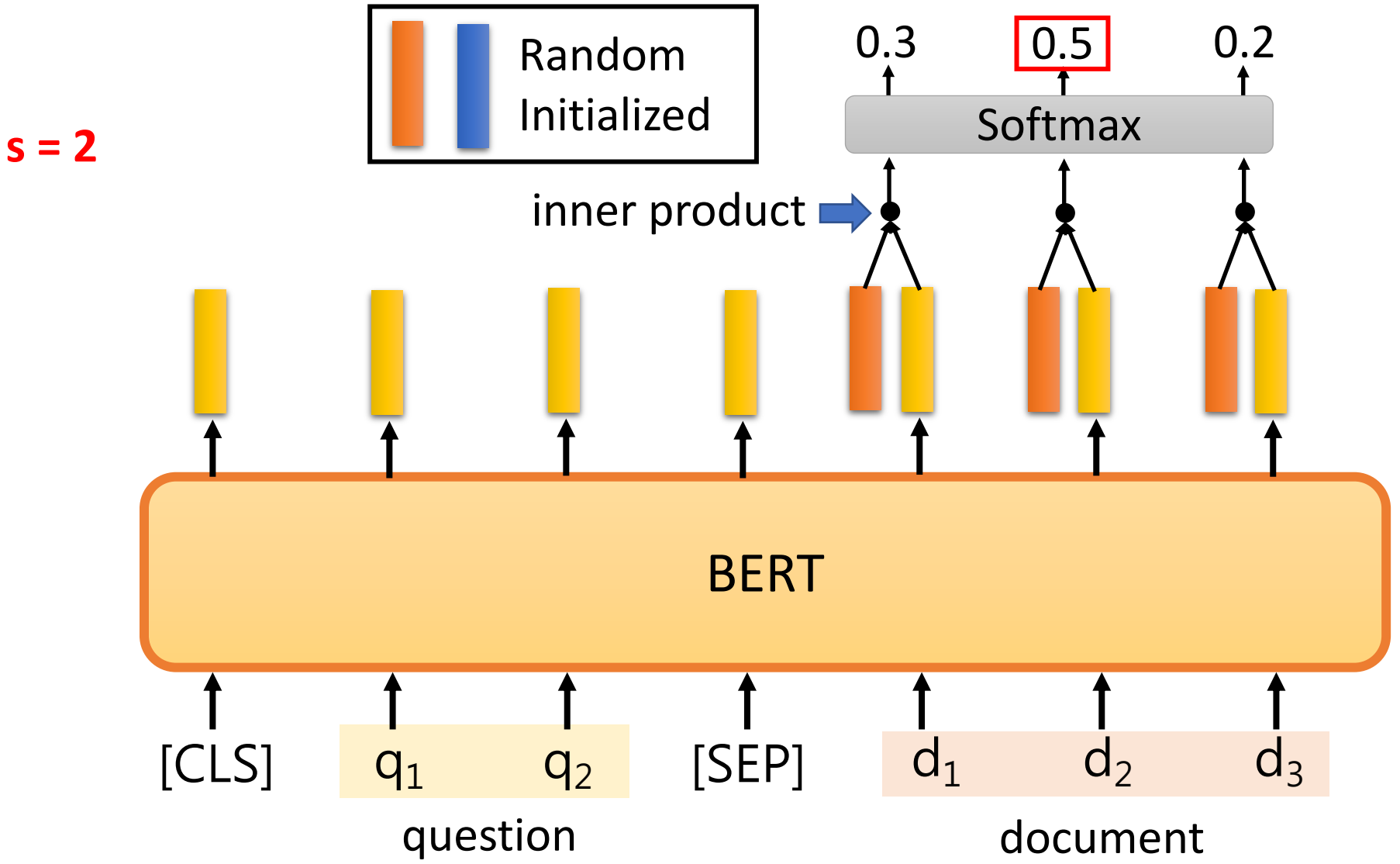
In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation includ 17 le, rain, sleet, snow, graupel and hail... pitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain 77 atte 79 cations are called "showers".

What causes precipitation to fall?
gravity $s = 17, e = 17$

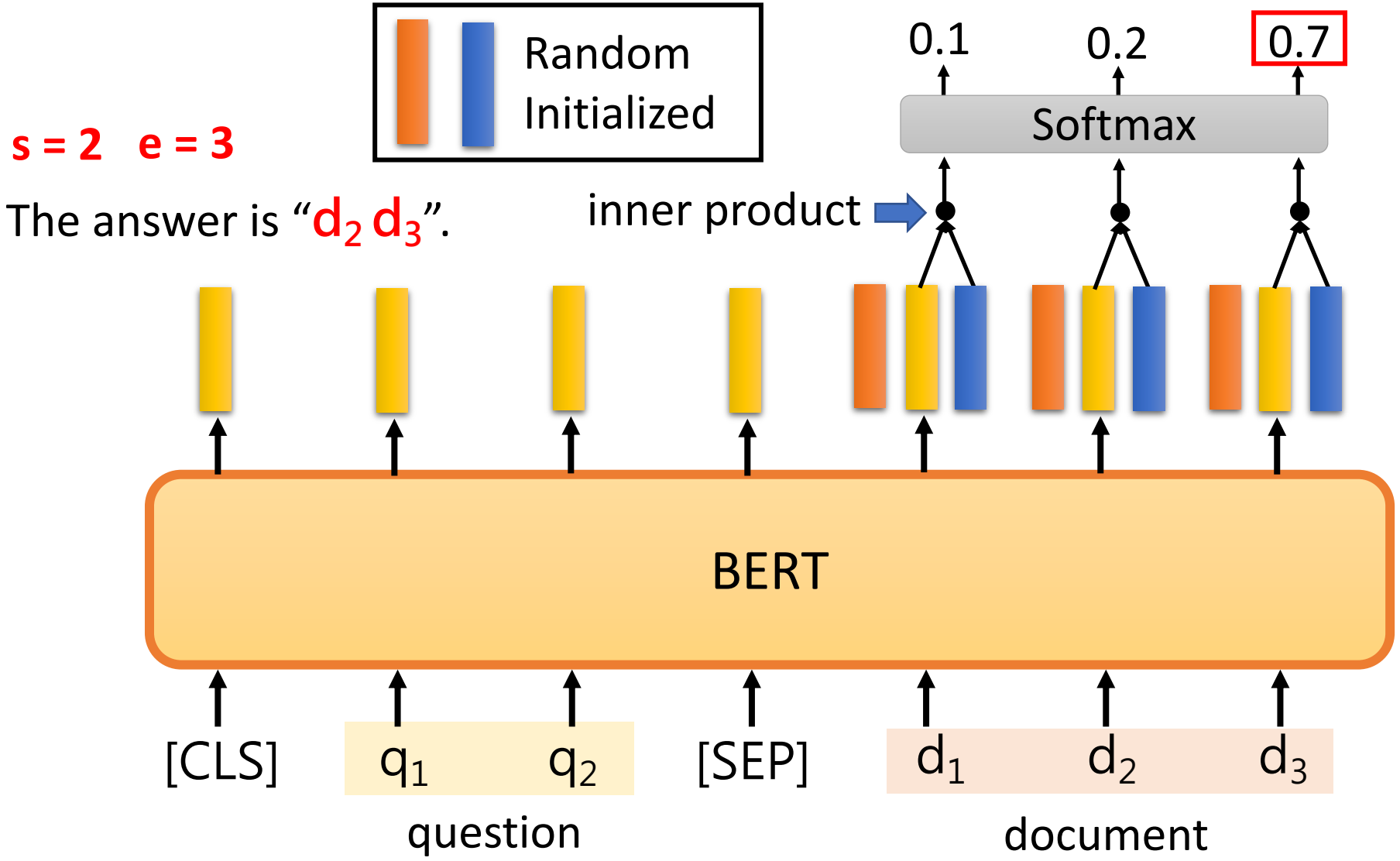
What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud $s = 77, e = 79$

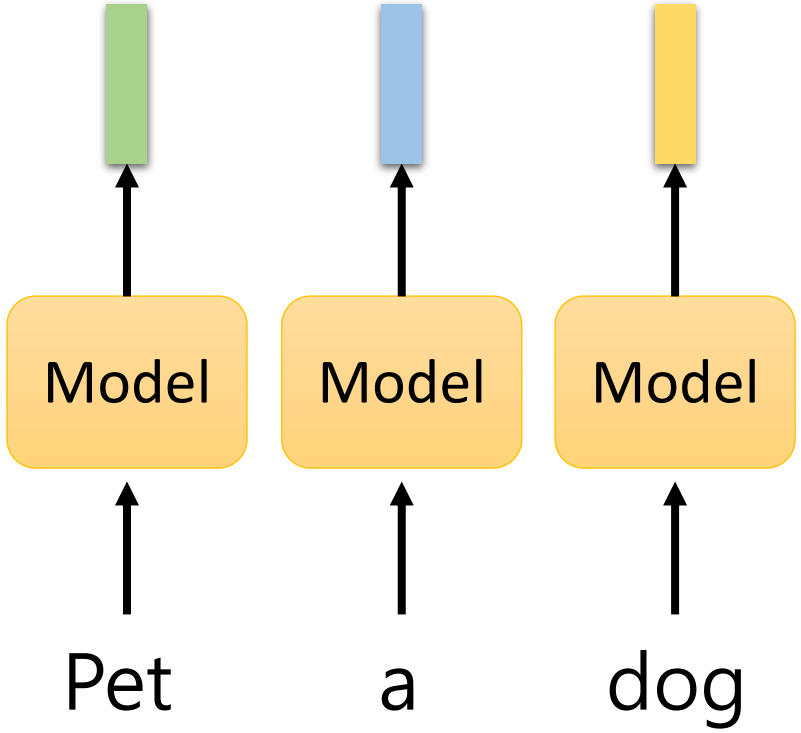
How to use BERT – Case 4



How to use BERT – Case 4

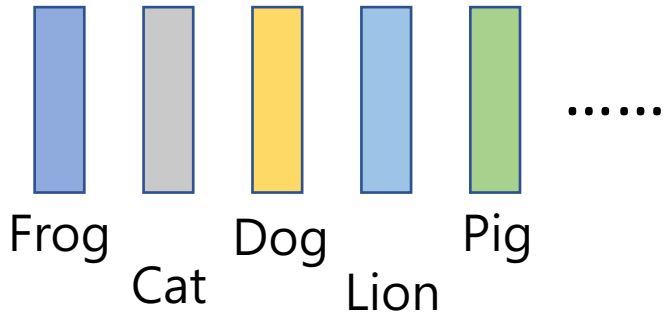


Contextualized word embedding



The token with the same type has the similar embedding.

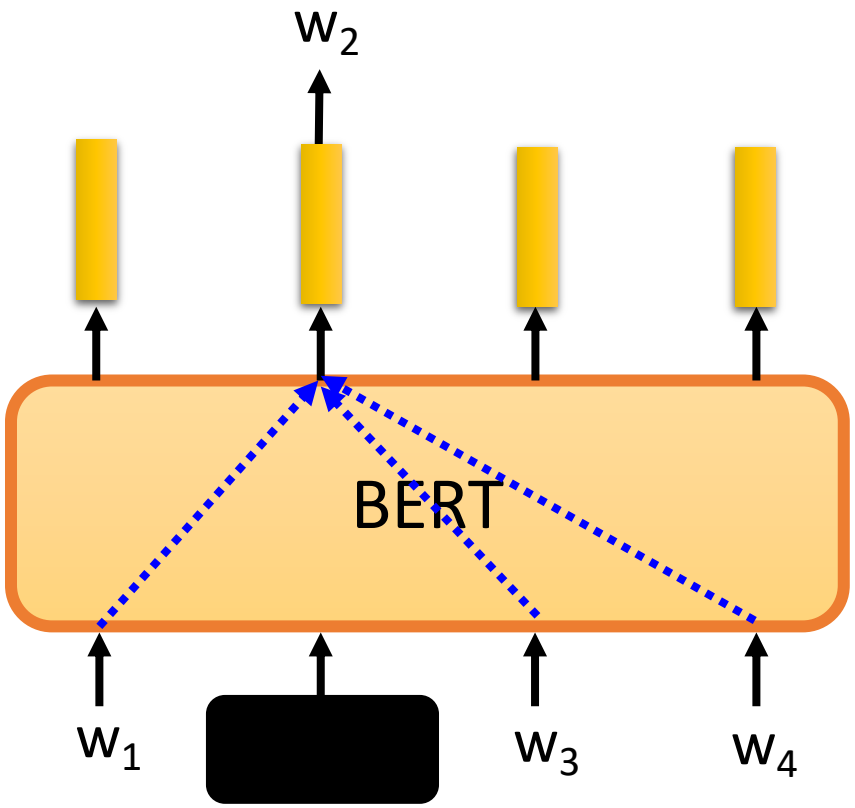
Simply a table look-up

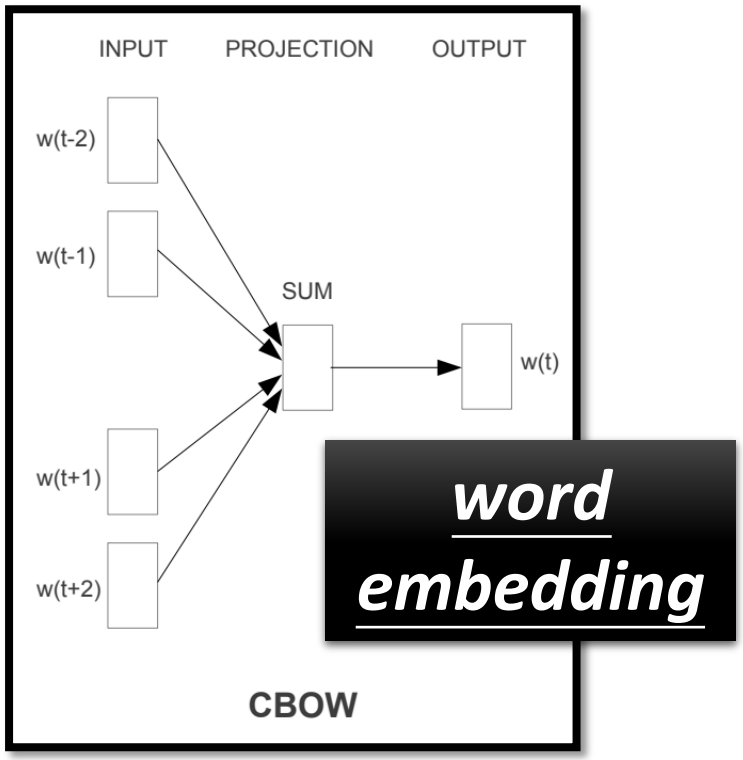


Word2vec [Mikolov, et al., NIPS'13]

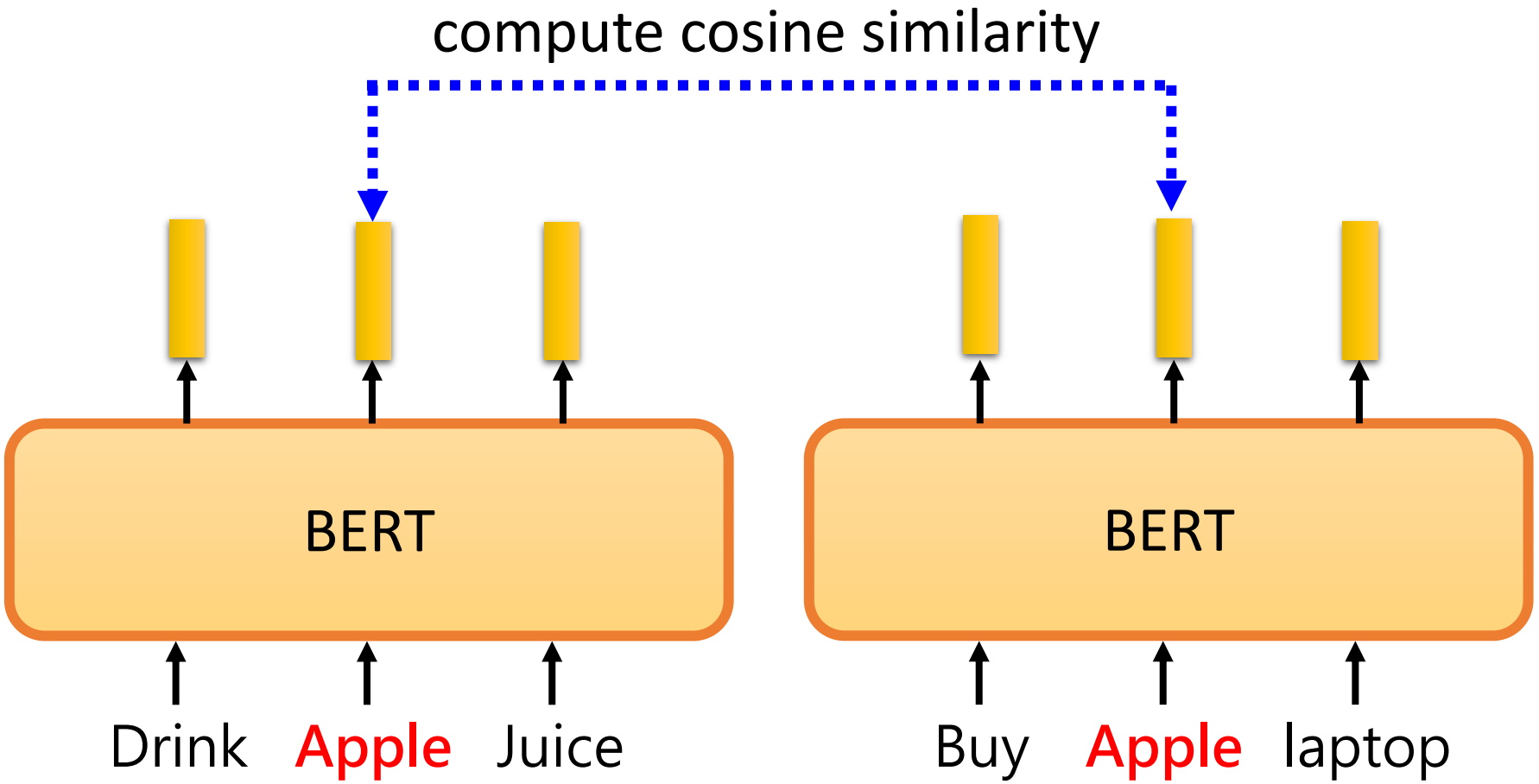
Glove [Pennington, et al., EMNLP'14]

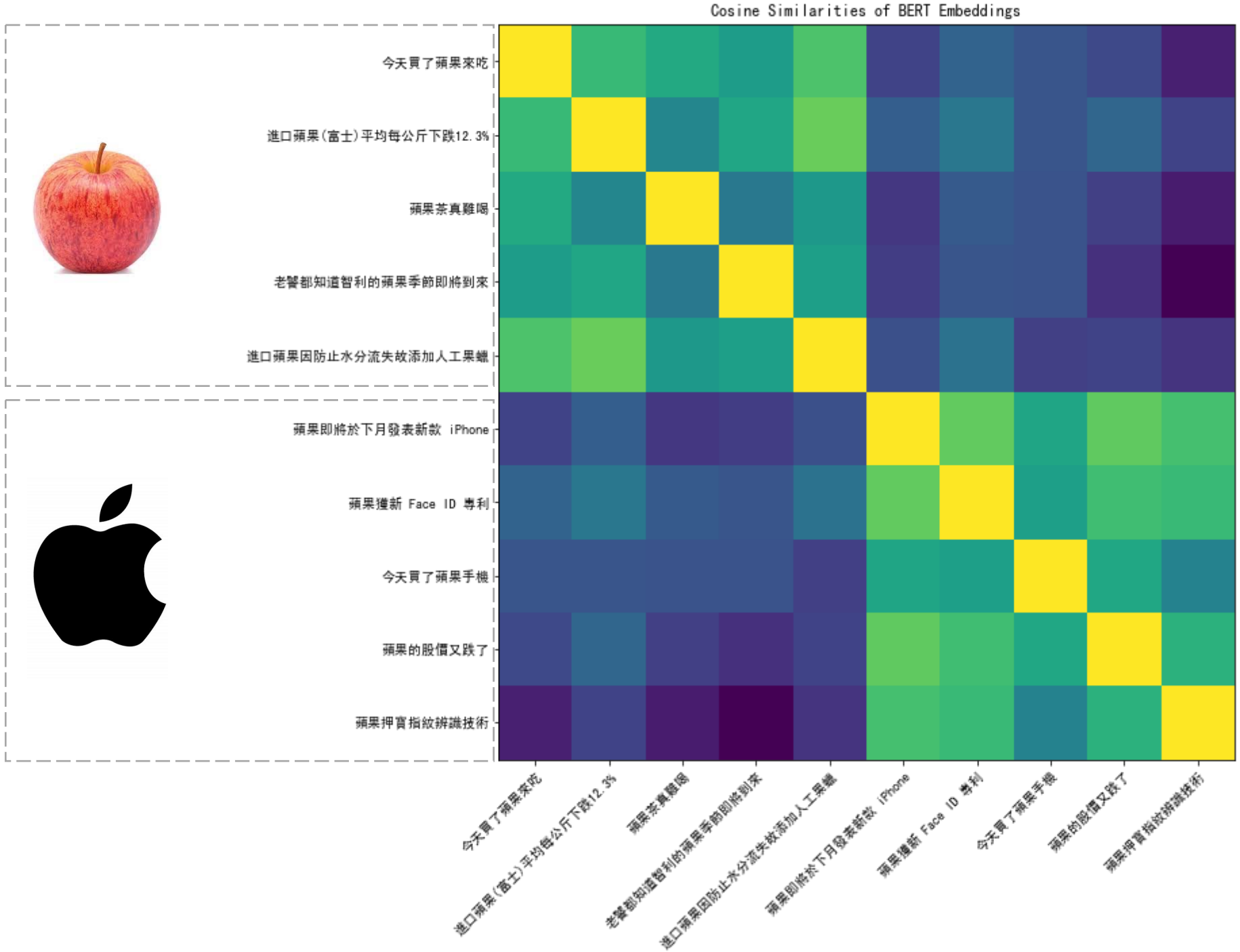
Contextualized word embedding





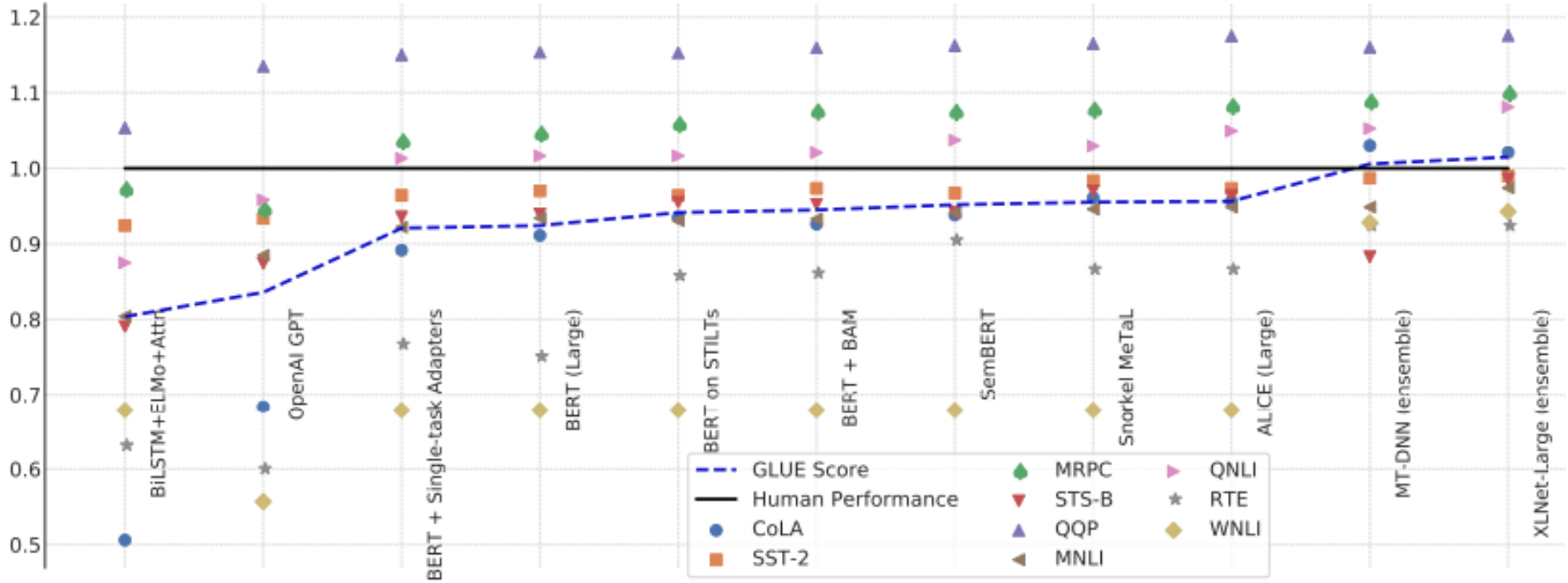
Contextualized word embedding




Cosine Similarities of BERT Embeddings
今天買了蘋果來吃
進口蘋果(富士)平均每公斤下跌12.3%
蘋果茶真難喝
老饕都知道智利的蘋果季節即將到來
進口蘋果因防止水分流失故添加人工果蠟
蘋果即將於下月發表新款 iPhone
蘋果獲新 Face ID 專利
今天買了蘋果手機
蘋果的股價又跌了
蘋果押寶指紋辨識技術

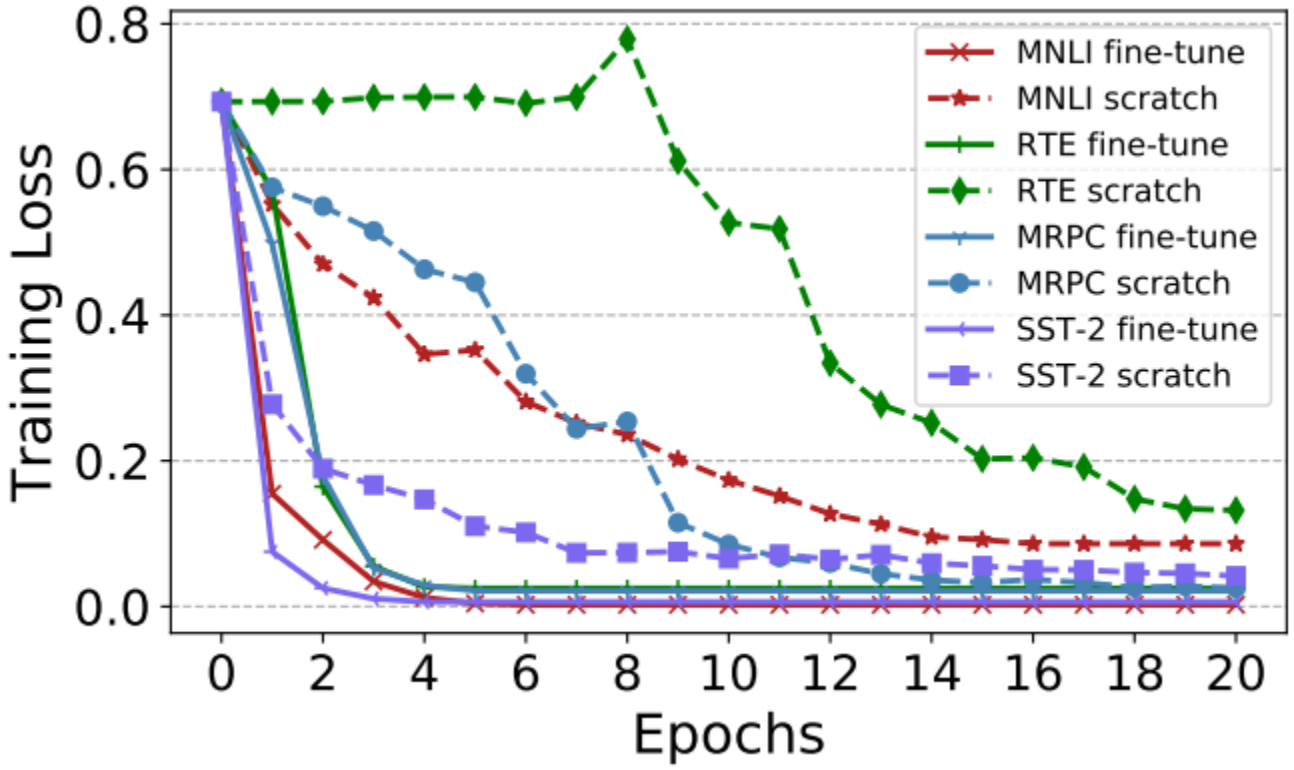
Why Pre-train Models?

- **GLUE**(General Language Understanding Evaluation) **scores**



Source of image: https://arxiv.org/abs/1905.00537

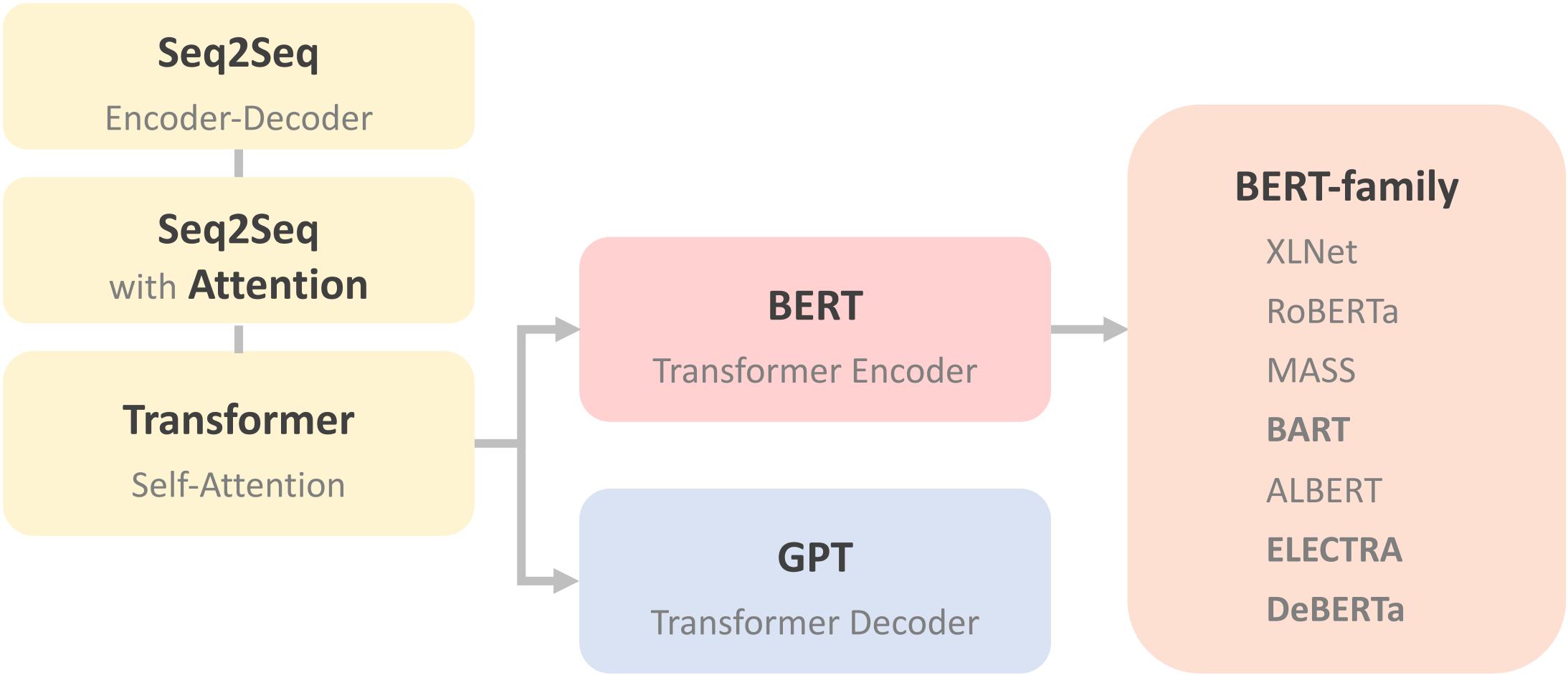
Why Fine-tune? (Pre-train v.s. Random)



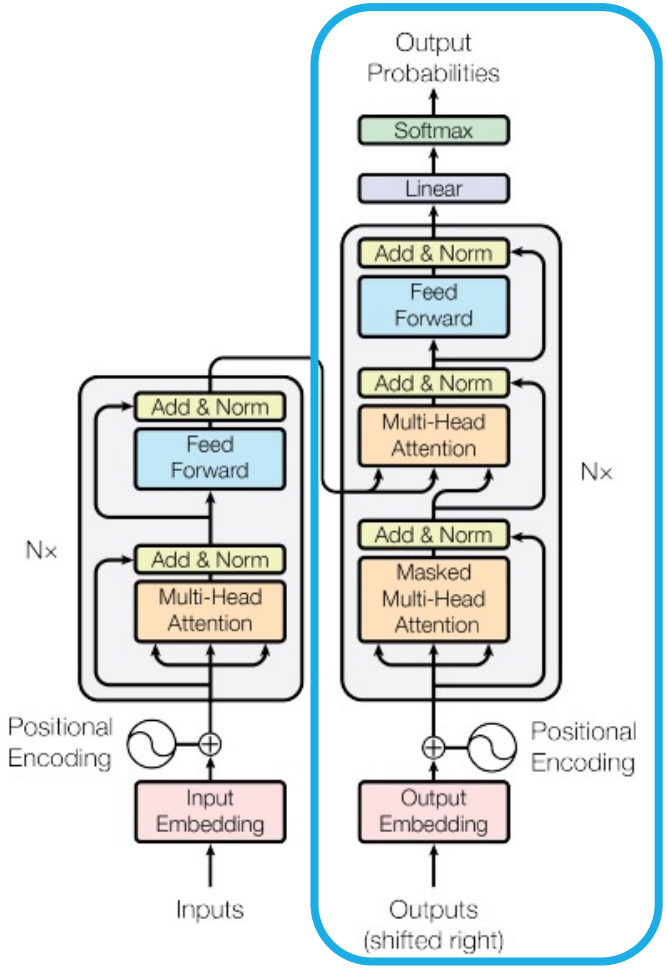
Source of image: https://arxiv.org/abs/1908.05620

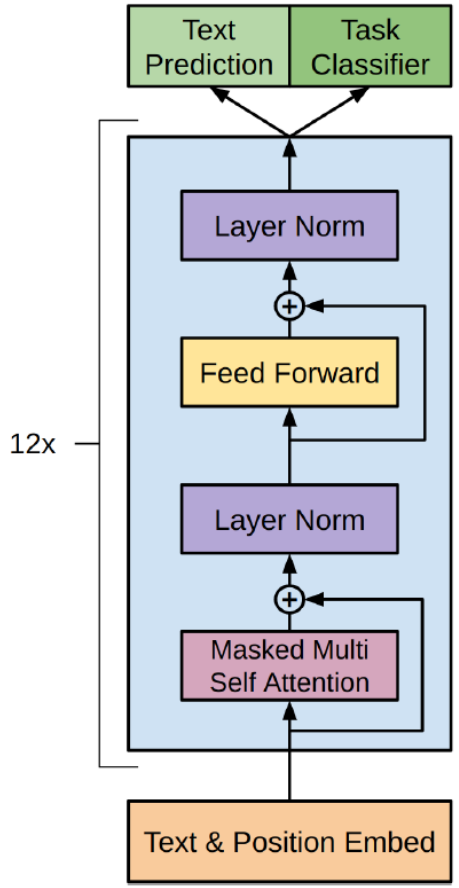
Various Pre-trained Models

Various Pre-trained Models



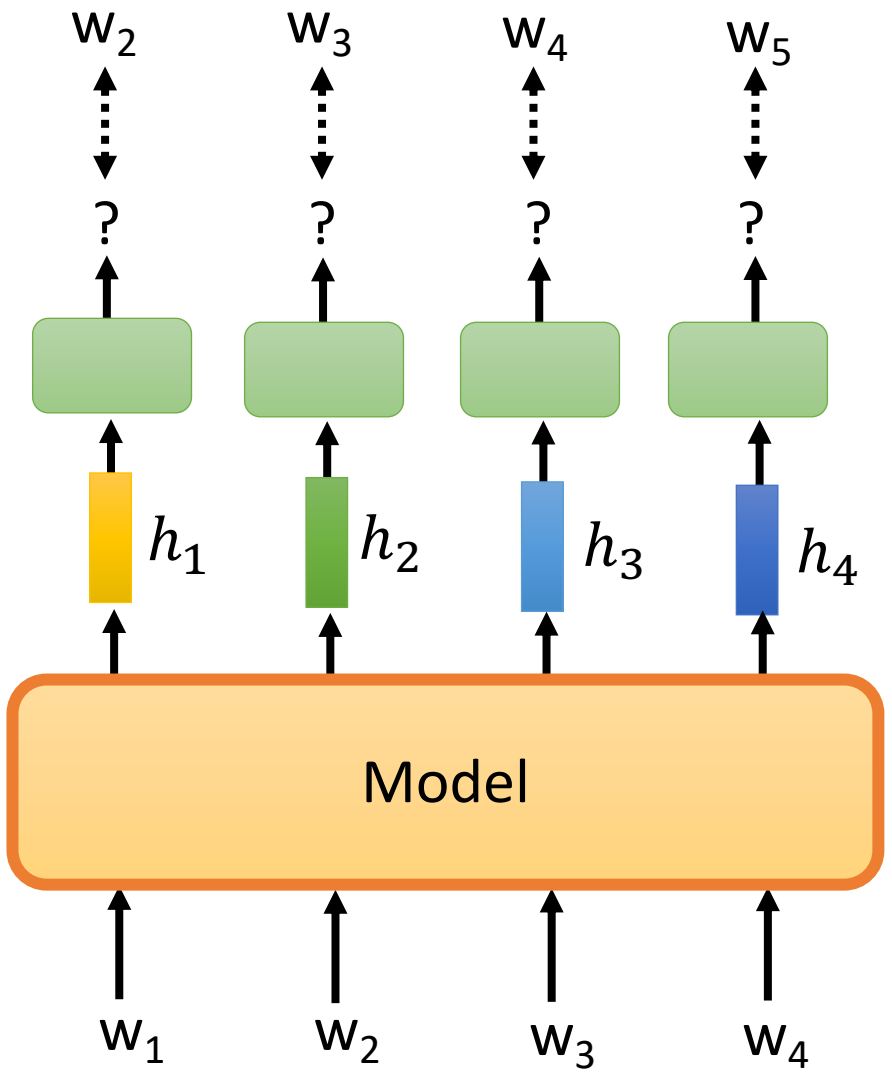
GENERATIVE PRE-TRAINED TRANSFORMER



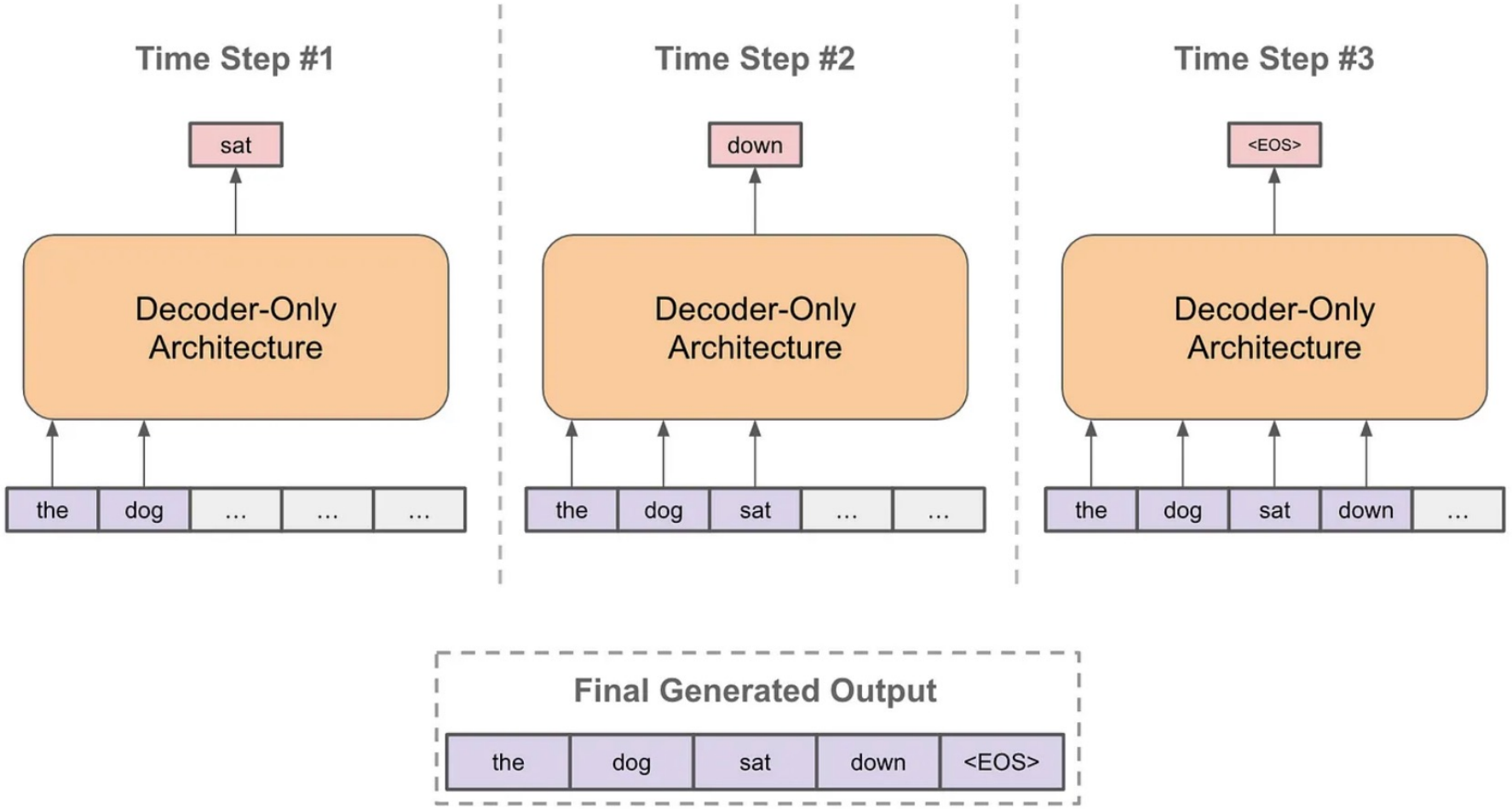


Transformer Decoder

Pre-train: Predict Next Token

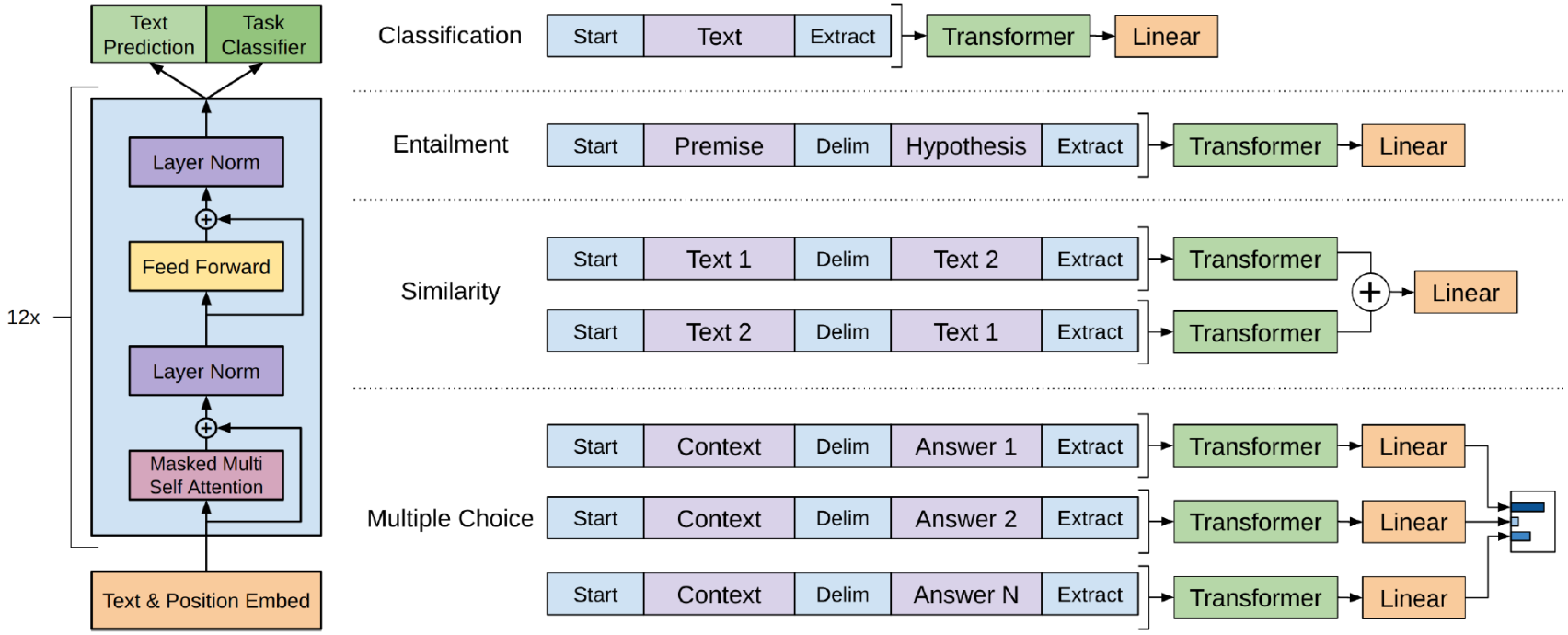


Pre-train: Predict Next Token



Fine-tuning

- The first generation of GPT is similar to other regular pre-trained model. Both are pre-trained using a self-supervised approach and then fine-tuned based on different downstream tasks using supervised learning. (GPT is good at text generation tasks)

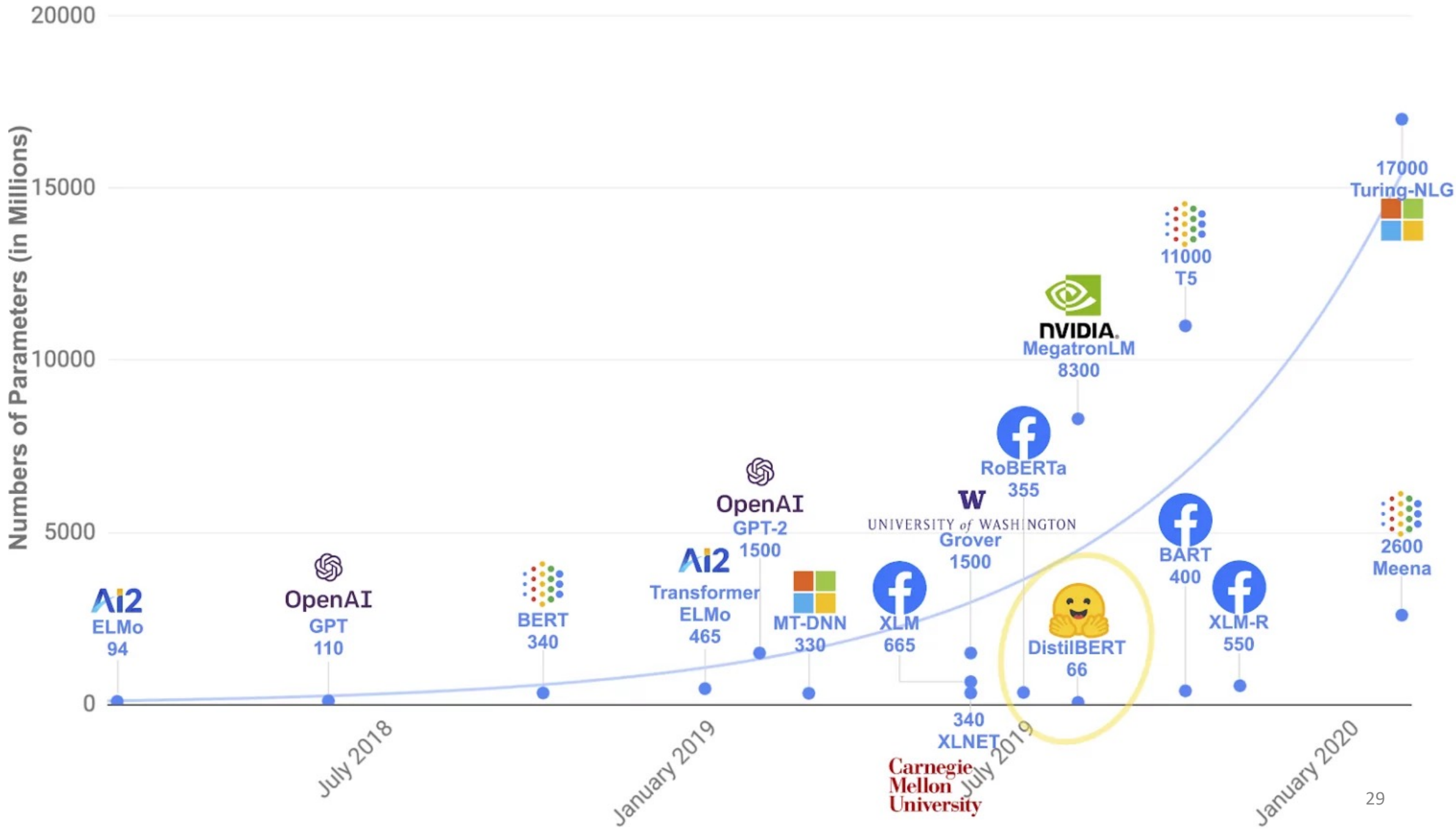


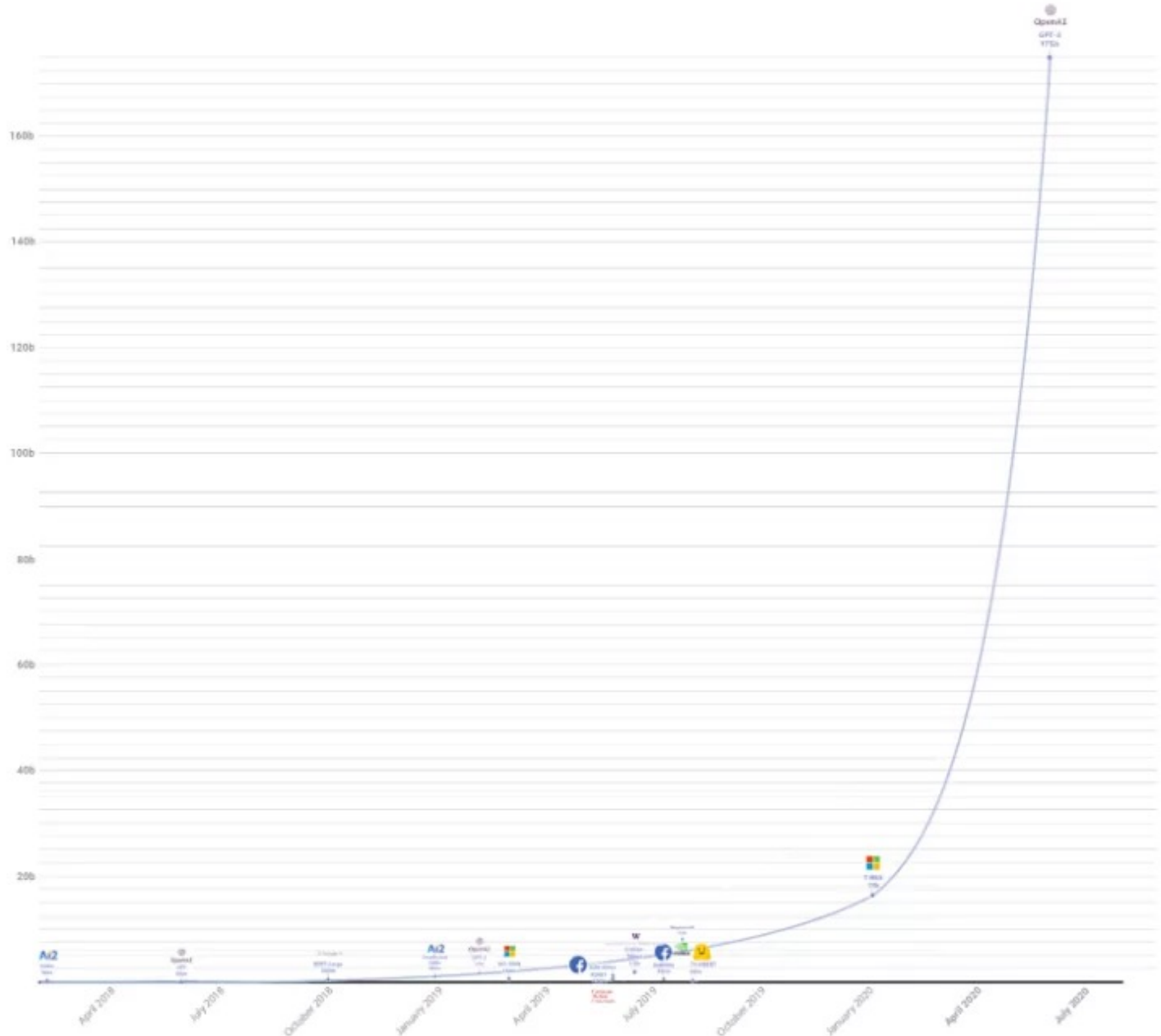
GPT-2 & gpt-3

The main difference between **GPT–2 & GPT–3** compared to GPT is that **fine–tuning** is no longer needed, instead, they utilize

1. Larger dataset for pre–training (40 GB, 45TB)
2. Larger model (1.542 billion, 175 billion)

to explore the performance of the model on each task.


20000
15000
10000
5000
0
Numbers of Parameters (in Millions)
ELMo
94
OpenAI
GPT
110
BERT
340
Transformer
ELMo
465
OpenAI
GPT-2
1500
MT-DNN
330
XLM
665
UNIVERSITY of WASHINGTON
Grover
1500
340
XLNET
Carnegie
Mellon
University
RoBERTa
355
NVIDIA
MegatronLM
8300
DistilBERT
66
11000
T5
BART
400
XLM-R
550
17000
Turing-NLG
2600
Meena
July 2018
January 2019
July 2019
January 2020


160b
140b
120b
100b
80b
60b
40b
20b
April 2018
July 2018
October 2018
January 2019
April 2019
July 2019
October 2019
January 2020
April 2020
July 2020

In-context Learning

"Few-shot"

(No GD Involved)

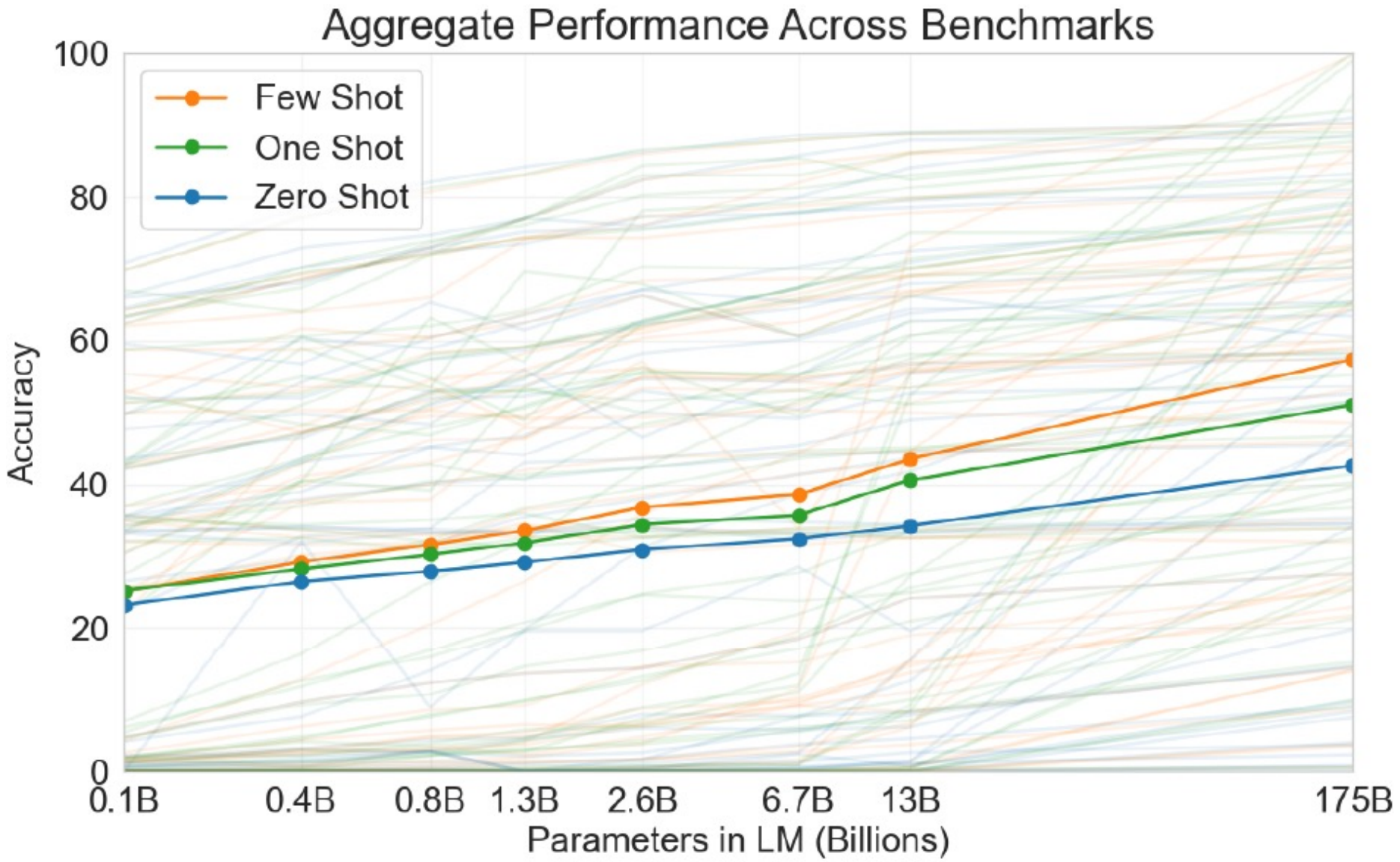
```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← examples
3  peppermint => menthe poivrée        ← examples
4  plush girafe => girafe peluche      ← examples
5  cheese =>                           ← prompt
```

"One-shot"

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← example
3  cheese =>                           ← prompt
```

"Zero-shot"

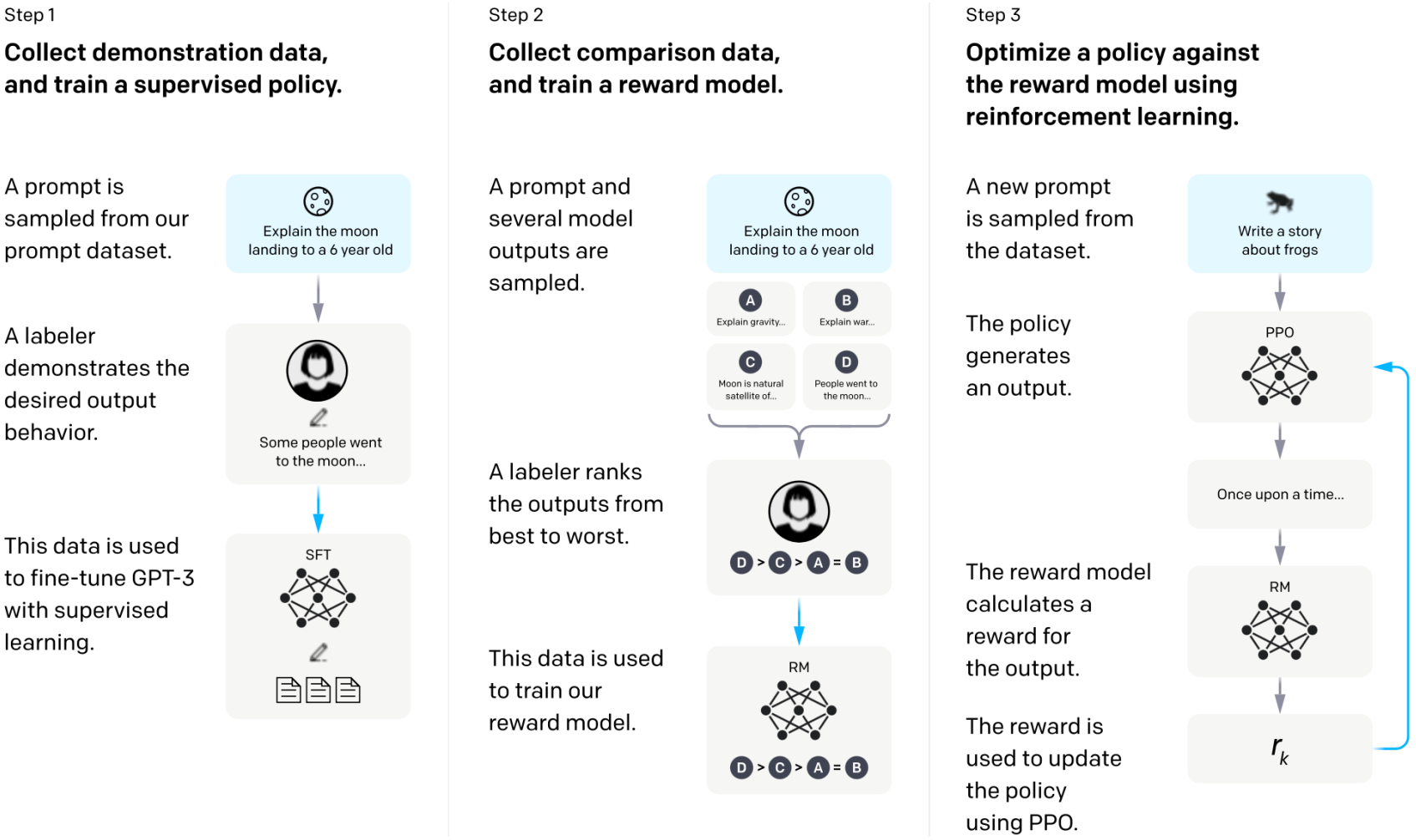
```
1  Translate English to French:        ← task description
2  cheese =>                           ← prompt
```



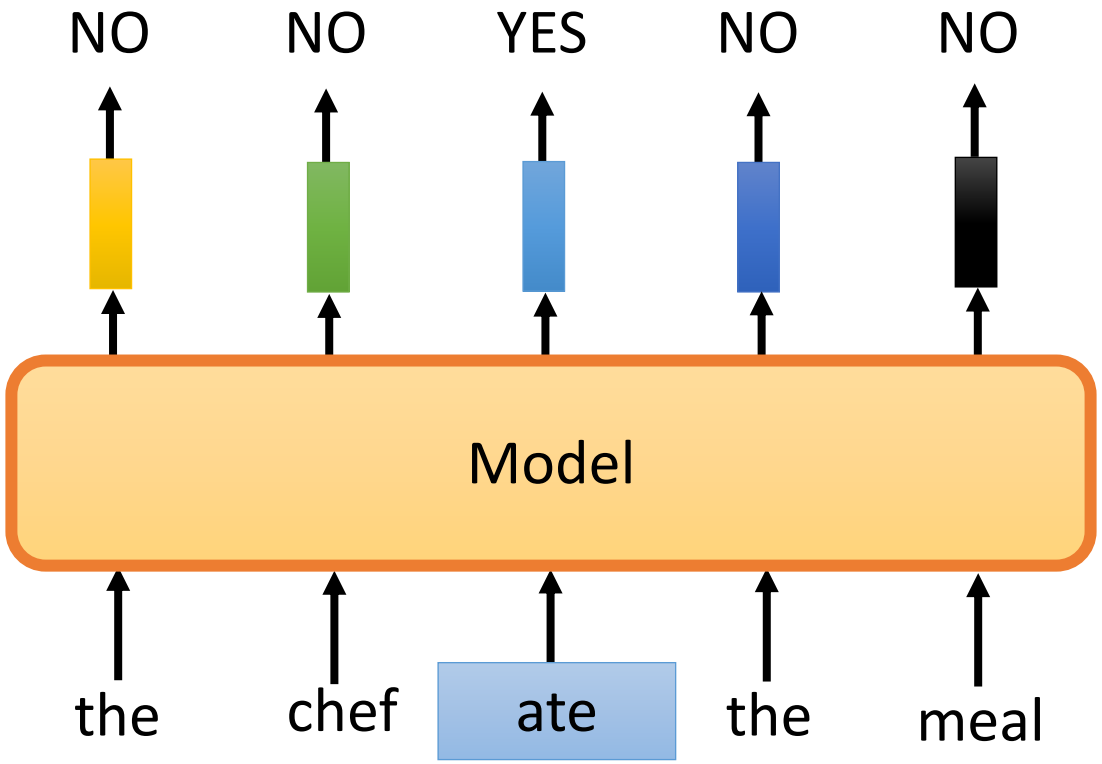
Average of **42** tasks

GPT-3.5、ChatGPT

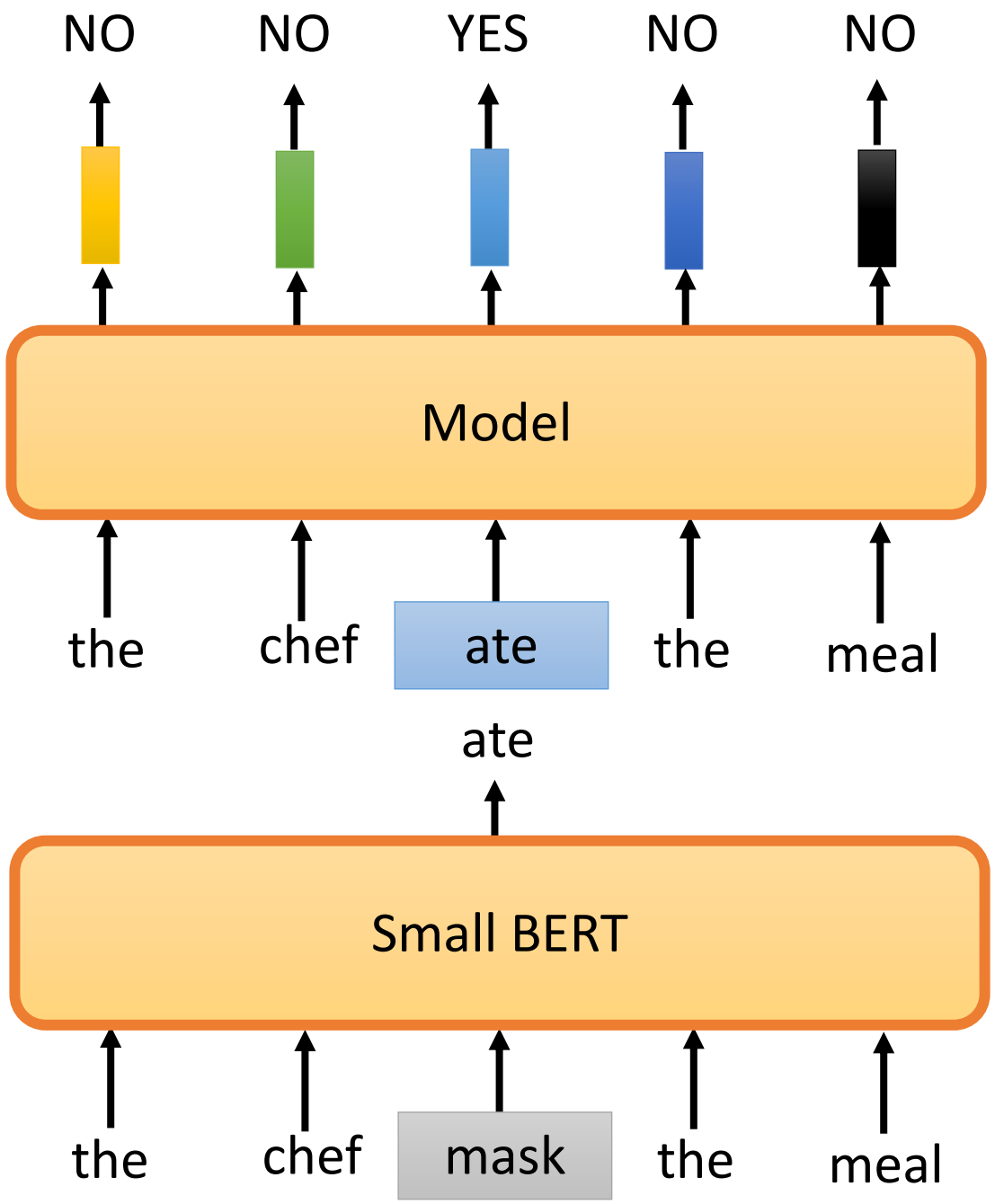
GPT–3.5 combines the models of GPT–3, Codex, InstructGPT, etc., and adds Reinforced Learning with Human Feedback (RLHF), which enhances the ability of program comprehension and generation, and has more realistic question and answer effects.

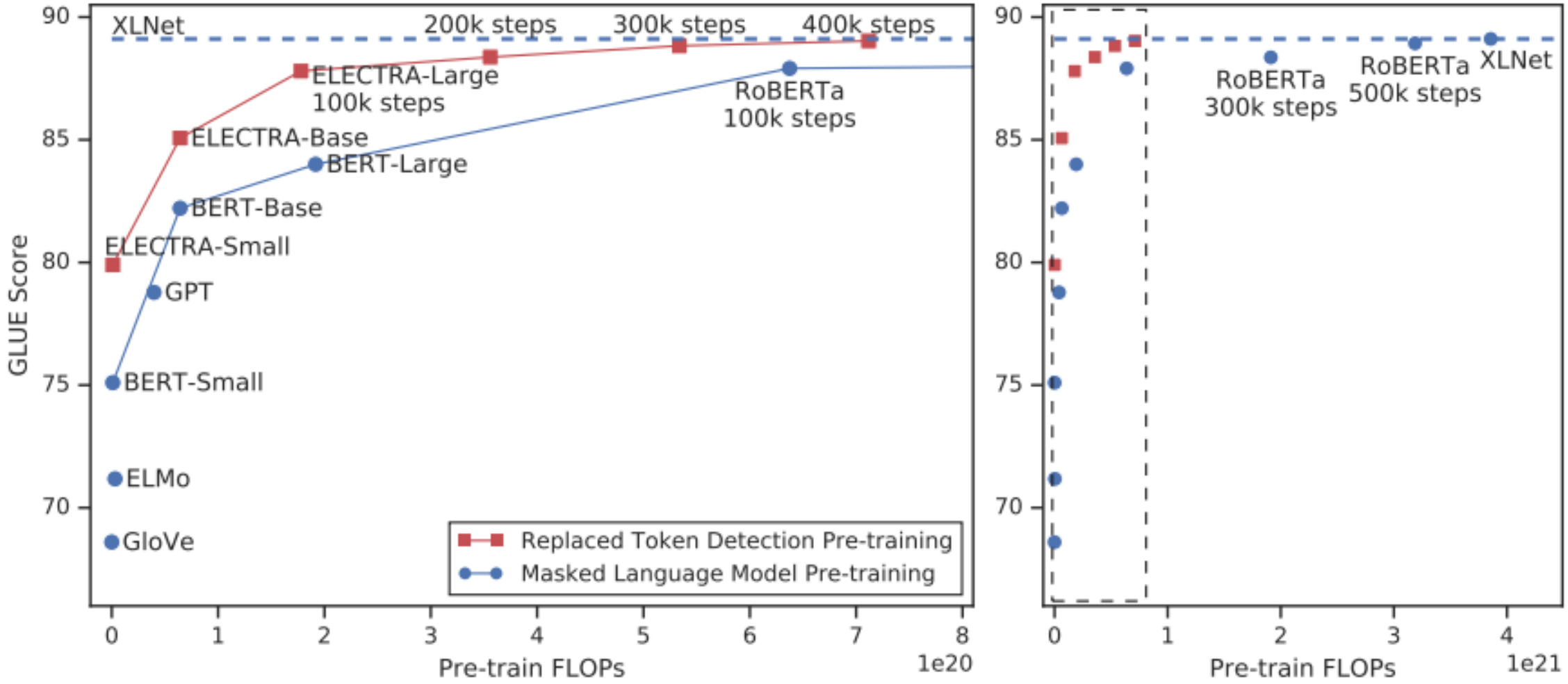


Efficiently Learning an Encoder that Classifies Token Replacements Accurately (ELECTRA)



Predicting yes/no is **easier** than reconstruction.


NO
NO
YES
NO
NO
Model
the
chef
ate
the
meal
ate
Small BERT
the
chef
mask
the
meal


XLNet
200k steps
300k steps
400k steps
ELECTRA-Large
100k steps
ELECTRA-Base
BERT-Large
BERT-Base
ELECTRA-Small
GPT
BERT-Small
ELMo
GloVe
RoBERTa
100k steps
RoBERTa
300k steps
RoBERTa
500k steps
XLNet
GLUE Score
Pre-train FLOPs
1e20
1e21
Replaced Token Detection Pre-training
Masked Language Model Pre-training

DeBERTa

Decoding-enhanced BERT **with disentangled attention**

This paper proposes two important approaches:

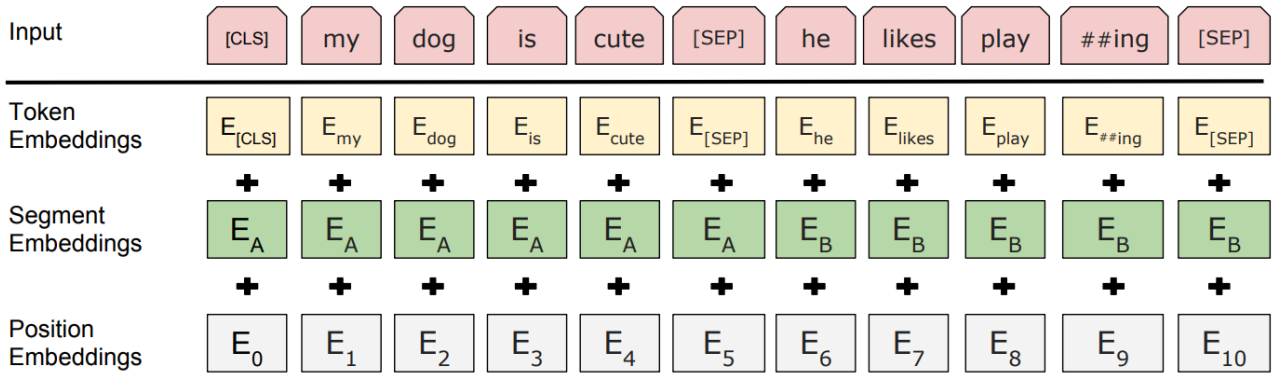
- **Disentangled Attention**
- **Enhanced Mask Decoder (EMD)**

Absolute Position Embedding

- Position Embedding in **Transformer** :

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$
$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

- Position Embedding in **BERT** :



Disentangled Attention

- In addition to the concept of Relative Position Embedding, DeBERTa proposes a new Attention Score design by separating Position Embedding from Content Embedding.

$$A_{i,j} = \{H_i, P_{i|j}\} \times \{H_j, P_{j|i}\}^\intercal$$
$$= H_i H_j^\intercal + H_i P_{j|i}^\intercal + P_{i|j} H_j^\intercal$$

(a) content-to-content (b) content-to-position (c) position-to-content

- Simple idea, the closer the words are, the higher the Attention Score should be.
- If "deep" is followed by "learning" in a sentence, the Attention Score should be higher than the separated two token.

Relative Distance in Disentangled Attention

- $\boldsymbol{k}$: maximum relative distance.
- $\delta(i,j) \in [0,2k)$: **relative** distance from token i to token j.

$$\delta(i,j) = \begin{cases} 0 & \text{for} \quad i-j \leqslant -k \\ 2k-1 & \text{for} \quad i-j \geqslant k \\ i-j+k & \text{others.} \end{cases}$$

[Example] $\boldsymbol{k = 3, i = 5}$

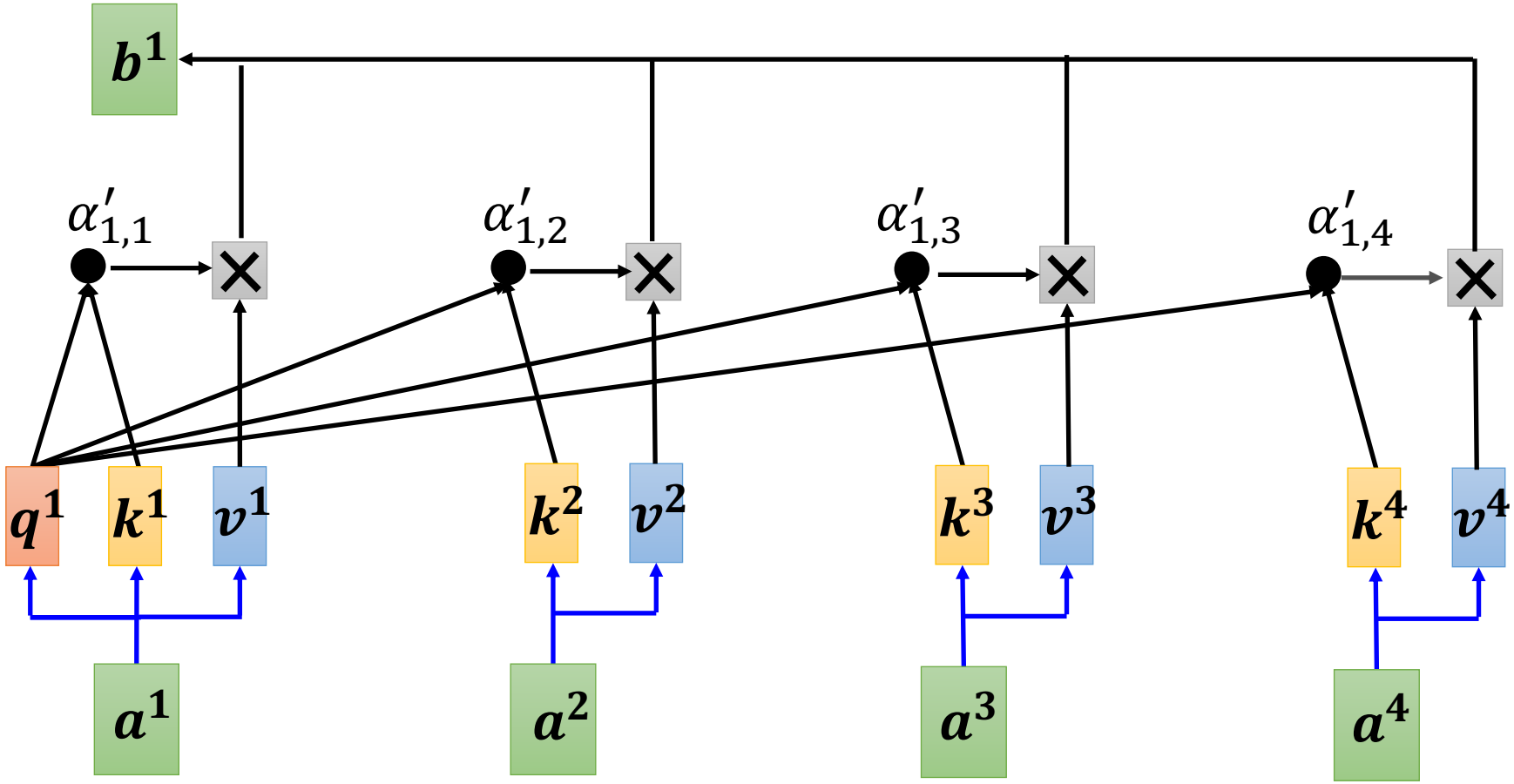
Index	1	2	3	4	5	6	7	8	9	10
$\boldsymbol{i-j}$	4	3	2	1	0	-1	-2	-3	-4	-5
$\boldsymbol{\delta(i,j)}$	6	6	5	4	3	2	1	0	0	0

$\boldsymbol{i = 5}$

Disentangled Attention

$$A_{i,j} = \{H_i, P_{i|j}\} \times \{H_j, P_{j|i}\}^\intercal$$
$$= \boxed{H_i H_j^\intercal} + H_i P_{j|i}^\intercal + P_{i|j} H_j^\intercal$$

(a) content-to-content (b) content-to-position (c) position-to-content

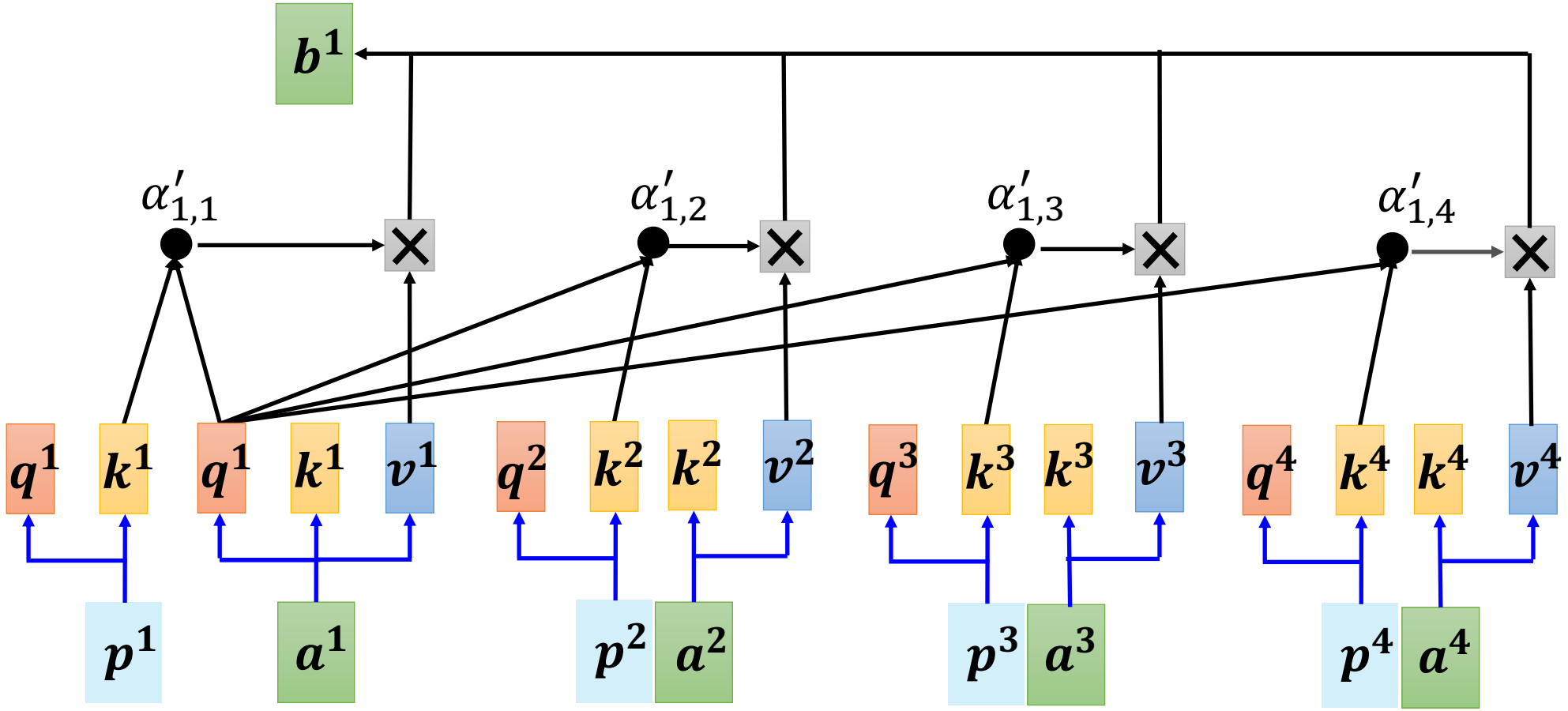


Content-to-content

Disentangled Attention

$$A_{i,j} = \{H_i, P_{i|j}\} \times \{H_j, P_{j|i}\}^\intercal$$
$$= H_i H_j^\intercal + H_i P_{j|i}^\intercal + P_{i|j} H_j^\intercal$$

(a) content-to-content (b) content-to-position (c) position-to-content

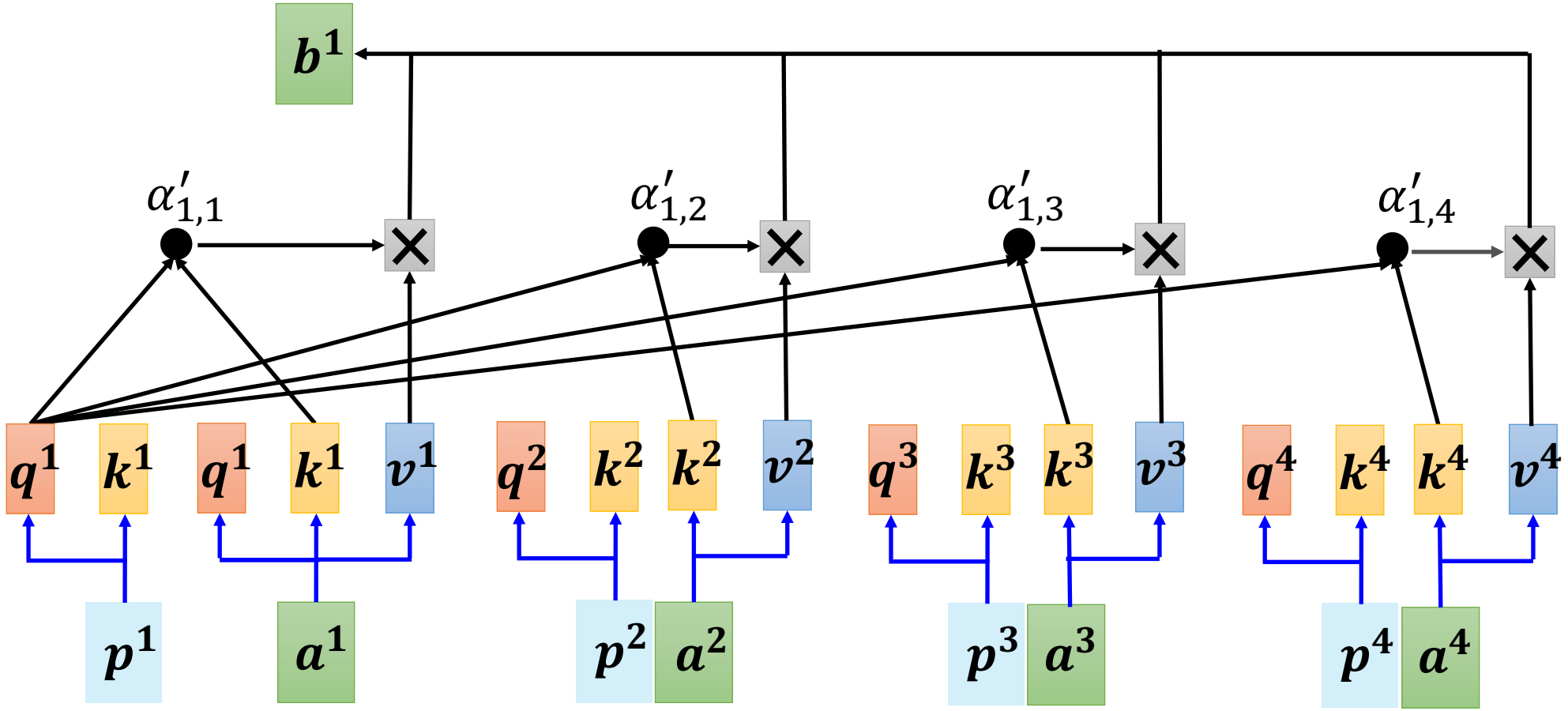


Content-to-position

Disentangled Attention

$$A_{i,j} = \{H_i, P_{i|j}\} \times \{H_j, P_{j|i}\}^\intercal$$
$$= H_i H_j^\intercal + H_i P_{j|i}^\intercal + P_{i|j} H_j^\intercal$$

(a) content-to-content (b) content-to-position (c) position-to-content



Position-to-content

Disentangled Attention

$$\boldsymbol{Q} = \boldsymbol{H}\boldsymbol{W}_{\boldsymbol{q}}, \boldsymbol{K} = \boldsymbol{H}\boldsymbol{W}_{\boldsymbol{k}}, \boldsymbol{V} = \boldsymbol{H}\boldsymbol{W}_{\boldsymbol{v}}, \boldsymbol{A} = \frac{\boldsymbol{Q}\boldsymbol{K}^{\intercal}}{\sqrt{d}}$$

$$\boldsymbol{H}_{\boldsymbol{o}} = \mathrm{softmax}(\boldsymbol{A})\boldsymbol{V}$$

Single-head Attention

$$\boldsymbol{Q}_{\boldsymbol{c}} = \boldsymbol{H}\boldsymbol{W}_{\boldsymbol{q},\boldsymbol{c}}, \boldsymbol{K}_{\boldsymbol{c}} = \boldsymbol{H}\boldsymbol{W}_{\boldsymbol{k},\boldsymbol{c}}, \boldsymbol{V}_{\boldsymbol{c}} = \boldsymbol{H}\boldsymbol{W}_{\boldsymbol{v},\boldsymbol{c}}, \boldsymbol{Q}_{\boldsymbol{r}} = \boldsymbol{P}\boldsymbol{W}_{\boldsymbol{q},\boldsymbol{r}}, \boldsymbol{K}_{\boldsymbol{r}} = \boldsymbol{P}\boldsymbol{W}_{\boldsymbol{k},\boldsymbol{r}}$$

$$\tilde{A}_{i,j} = \underbrace{\boldsymbol{Q}_{\boldsymbol{i}}^{\boldsymbol{c}}\boldsymbol{K}_{\boldsymbol{j}}^{\boldsymbol{c}\intercal}}_{\text{(a) content-to-content}} + \underbrace{\boldsymbol{Q}_{\boldsymbol{i}}^{\boldsymbol{c}}\boldsymbol{K}_{\boldsymbol{\delta}(\boldsymbol{i},\boldsymbol{j})}^{\boldsymbol{r}\intercal}}_{\text{(b) content-to-position}} + \underbrace{\boldsymbol{K}_{\boldsymbol{j}}^{\boldsymbol{c}}\boldsymbol{Q}_{\boldsymbol{\delta}(\boldsymbol{j},\boldsymbol{i})}^{\boldsymbol{r}\intercal}}_{\text{(c) position-to-content}}$$

$$\boldsymbol{H}_{\boldsymbol{o}} = \mathrm{softmax}(\frac{\tilde{\boldsymbol{A}}}{\sqrt{3d}})\boldsymbol{V}_{\boldsymbol{c}}$$

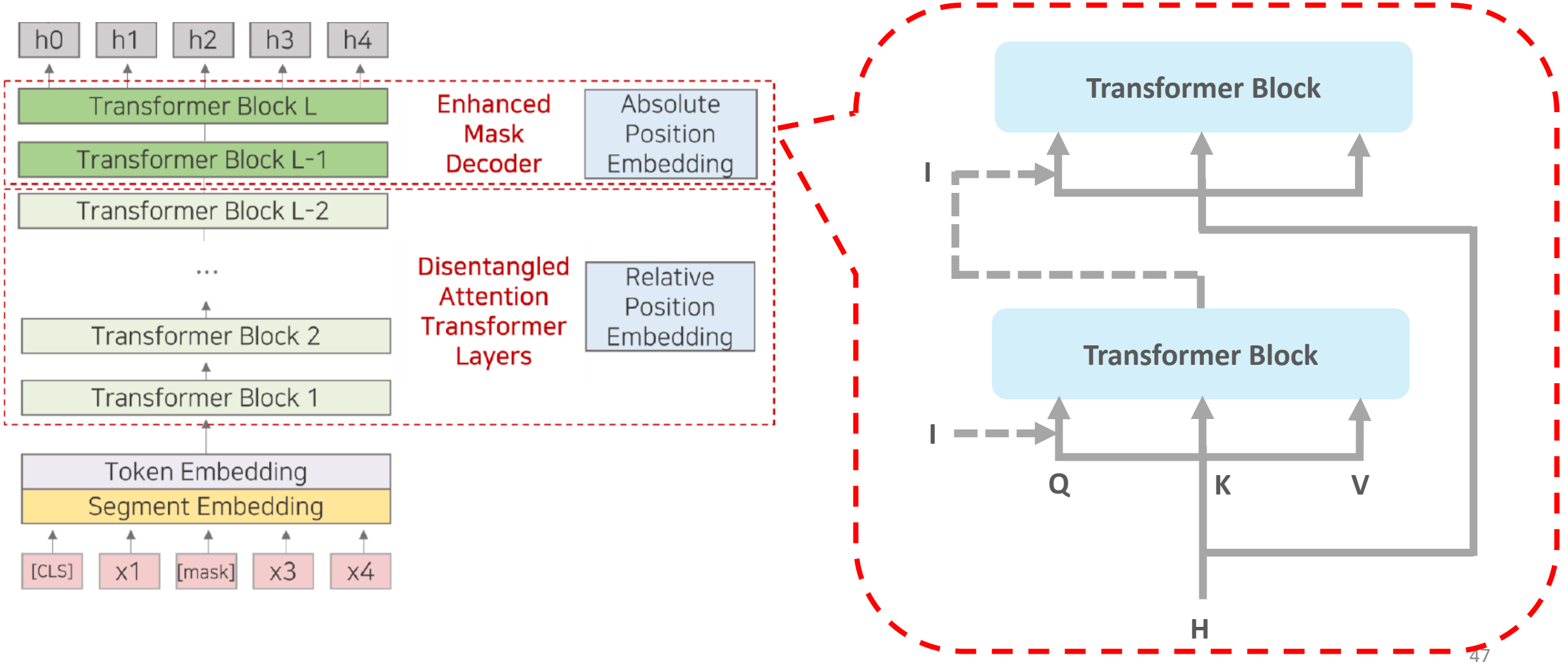
Disentangled Attention

Enhanced Mask Decoder(EMD)

Disentangled Attention didn’t consider the information from Absolute Position Embedding. So, EMD would add **Absolute Position Embedding** as extra information before decoder

a new store opened beside the new mall

Enhanced Mask Decoder(EMD)



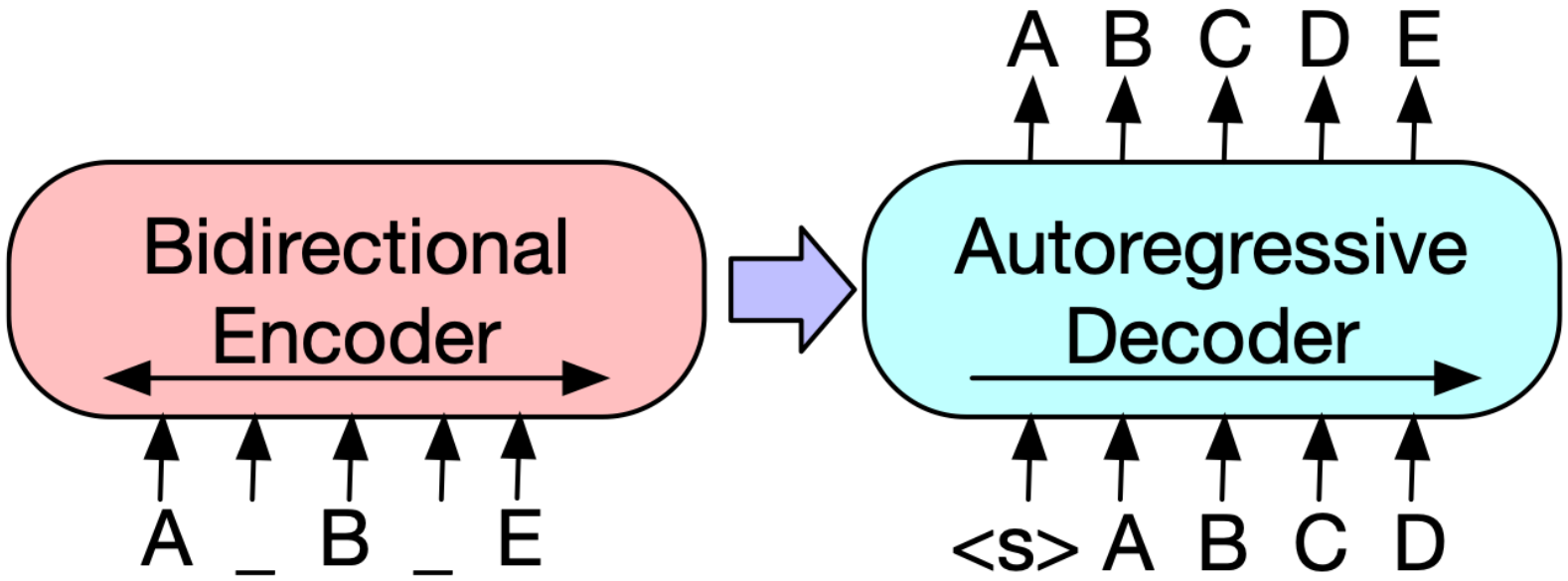
DeBERTa

Model	CoLA Mcc	QQP Acc	MNLI-m/mm Acc	SST-2 Acc	STS-B Corr	QNLI Acc	RTE Acc	MRPC Acc	Avg.
BERT_{large}	60.6	91.3	86.6/-	93.2	90.0	92.3	70.4	88.0	84.05
RoBERTa_{large}	68.0	92.2	90.2/90.2	96.4	92.4	93.9	86.6	90.9	88.82
XLNet_{large}	69.0	92.3	90.8/90.8	97.0	92.5	94.9	85.9	90.8	89.15
$\text{ELECTRA}_{\text{large}}$	69.1	92.4	90.9/-	96.9	92.6	95.0	88.0	90.8	89.46
DeBERTa_{large}	70.5	92.3	91.1/91.1	96.8	92.8	95.3	88.3	91.9	90.00

Table 1: Comparison results on the GLUE development set.

Bart (Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension)

BART can be seen as generalizing BERT (due to the bidirectional encoder) and GPT2 (with the left to right decoder).

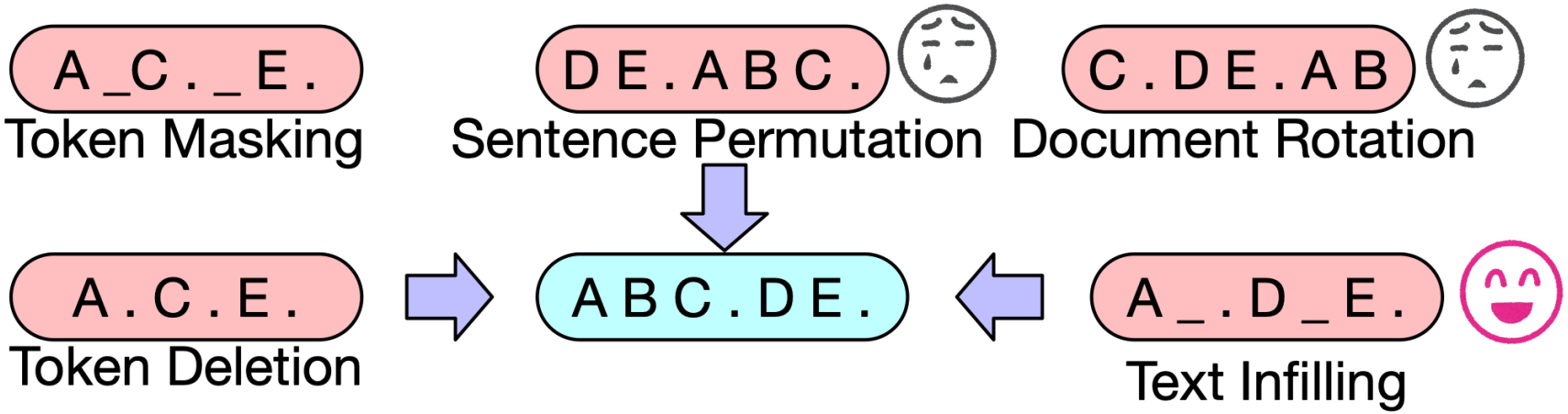


What BART improve?

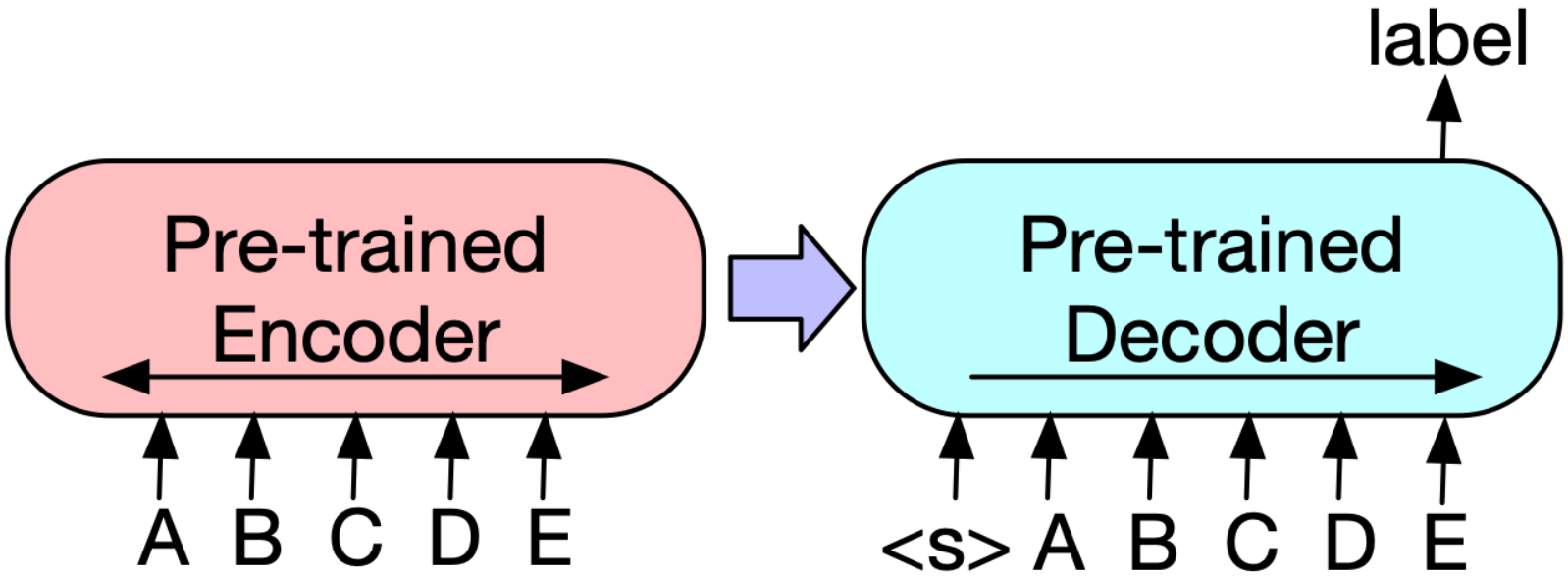
1. Added Decoder to BERT to make it suitable for seq2seq.

2. Replaced MLM tasks with more complex pre-training tasks.

PRETRAINING TASKS

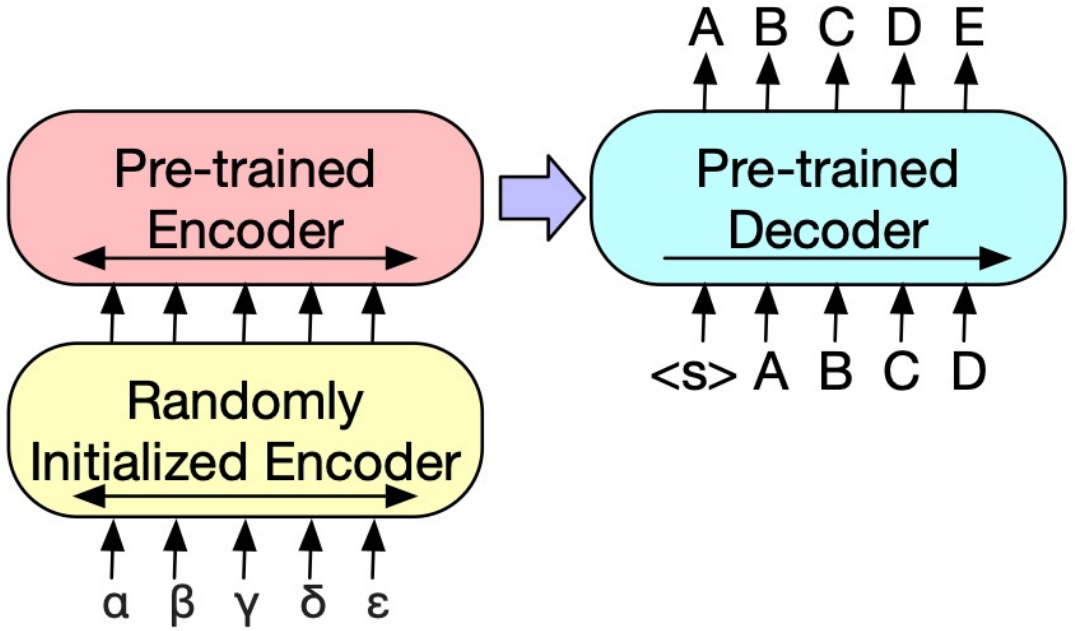


FINE-TUNING BART : CLASSIFICATION



(a) To use BART for classification problems, the same input is fed into the encoder and decoder, and the representation from the final output is used.

FINE-TUNING BART : TRANSLATION



(b) For machine translation, we learn a small additional encoder that replaces the word embeddings in BART. The new encoder can use a disjoint vocabulary.

CNN/DAILY MAIL ABSTRACTIVE SUMMARIZATION TASK

Model	Rouge2	Model Size	Pretraining
PT-Gen	17.28	22 M	None
TransformerABS	17.76	200M	None
BERTSumABS	19.39	220 M	Encoder
UniLM	20.3	340 M	Seq2Seq
T5-base	20.34	770 M	Seq2Seq
BART	**21.28**	**406 M**	**Seq2Seq**
T5-11B	21.55	11 B	Seq2Seq

QA