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**

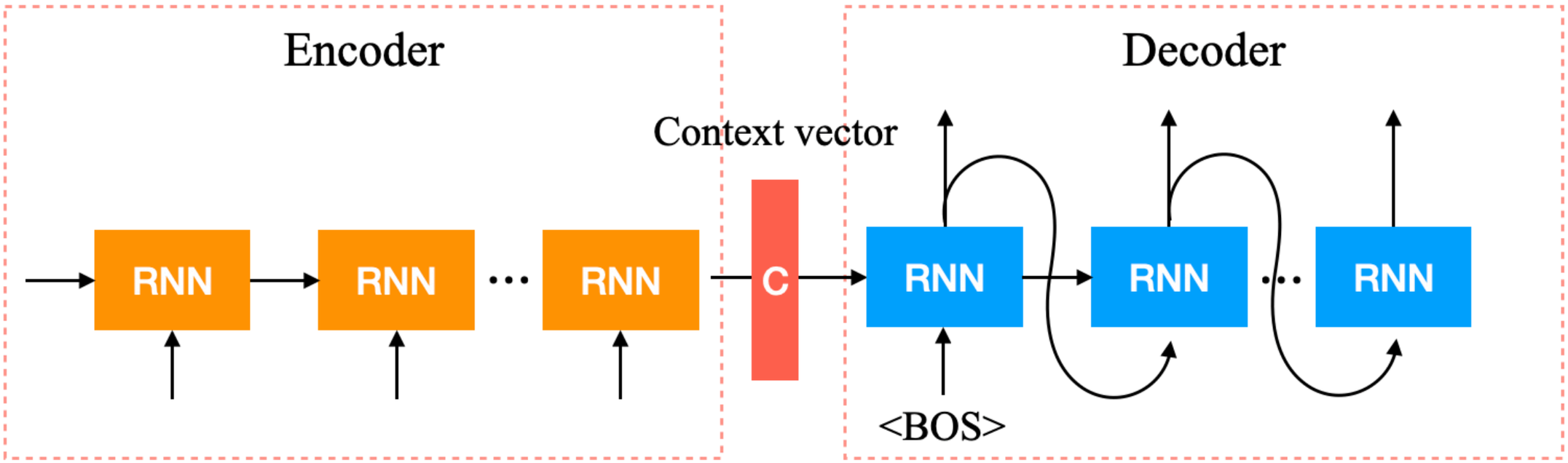
OpenAI

- Seq2seq
- Transformer
- Meta Learning
- Zero, One, Few Shot Learning
- GPT-3

Seq2seq

Seq2seq model:

能夠處理輸入與輸出不固定長度之序列任務



應用場景：

- Neural Machine Translation

輸入：Taiwan is a beautiful country.
輸出：台灣是一個美麗的國家。

- Chatbot

輸入：台灣是一個美麗的國家。
輸出：當然。

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." Advances in neural information processing systems. 2014.

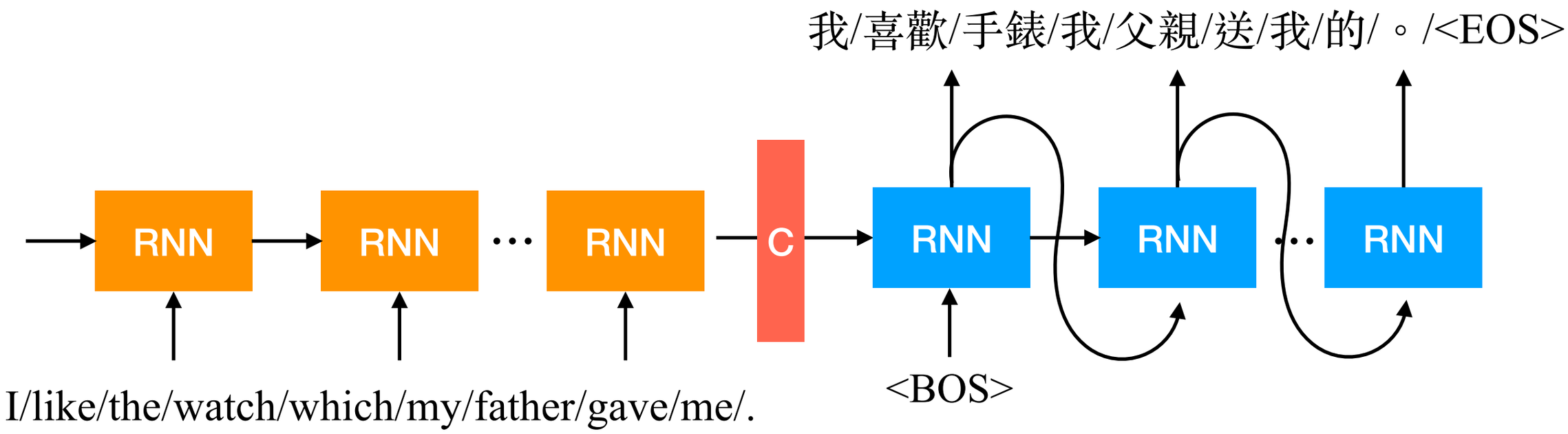
Seq2seq

輸入：I like the watch which my father gave me.

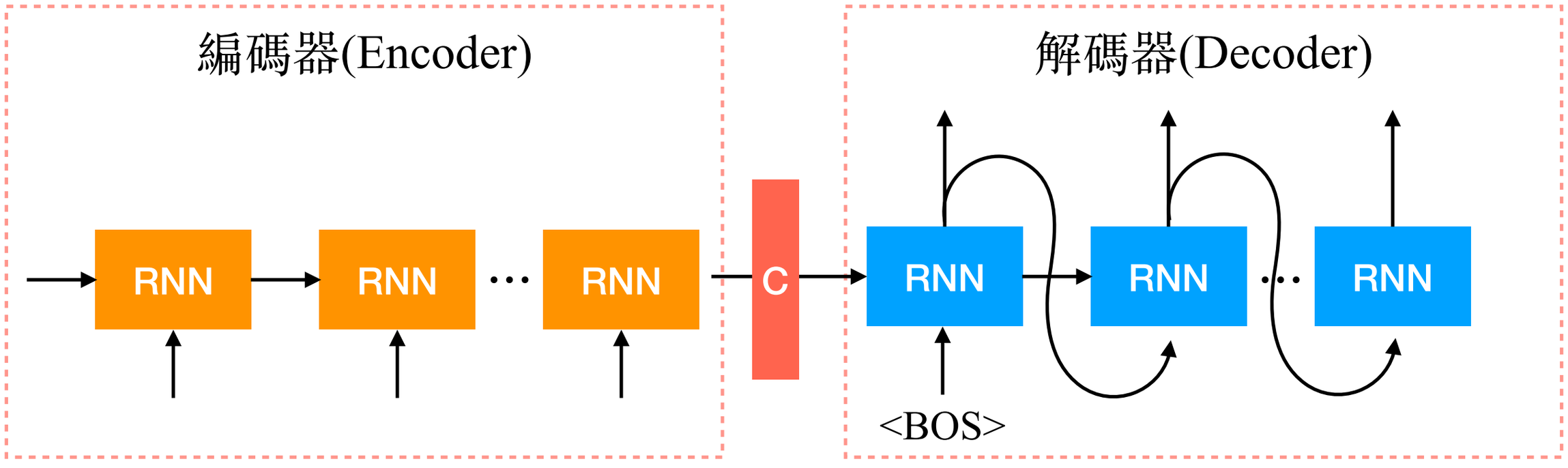
- Hard-Attention：我喜歡手錶我父親送我的。

輸出：

- Soft-Attention：我喜歡我父親送我的手錶。



Seq2seq



- **長期依賴(Long-term dependency)：**

當序列越來越長的時候，越前先的特徵會被RNN所遺忘。

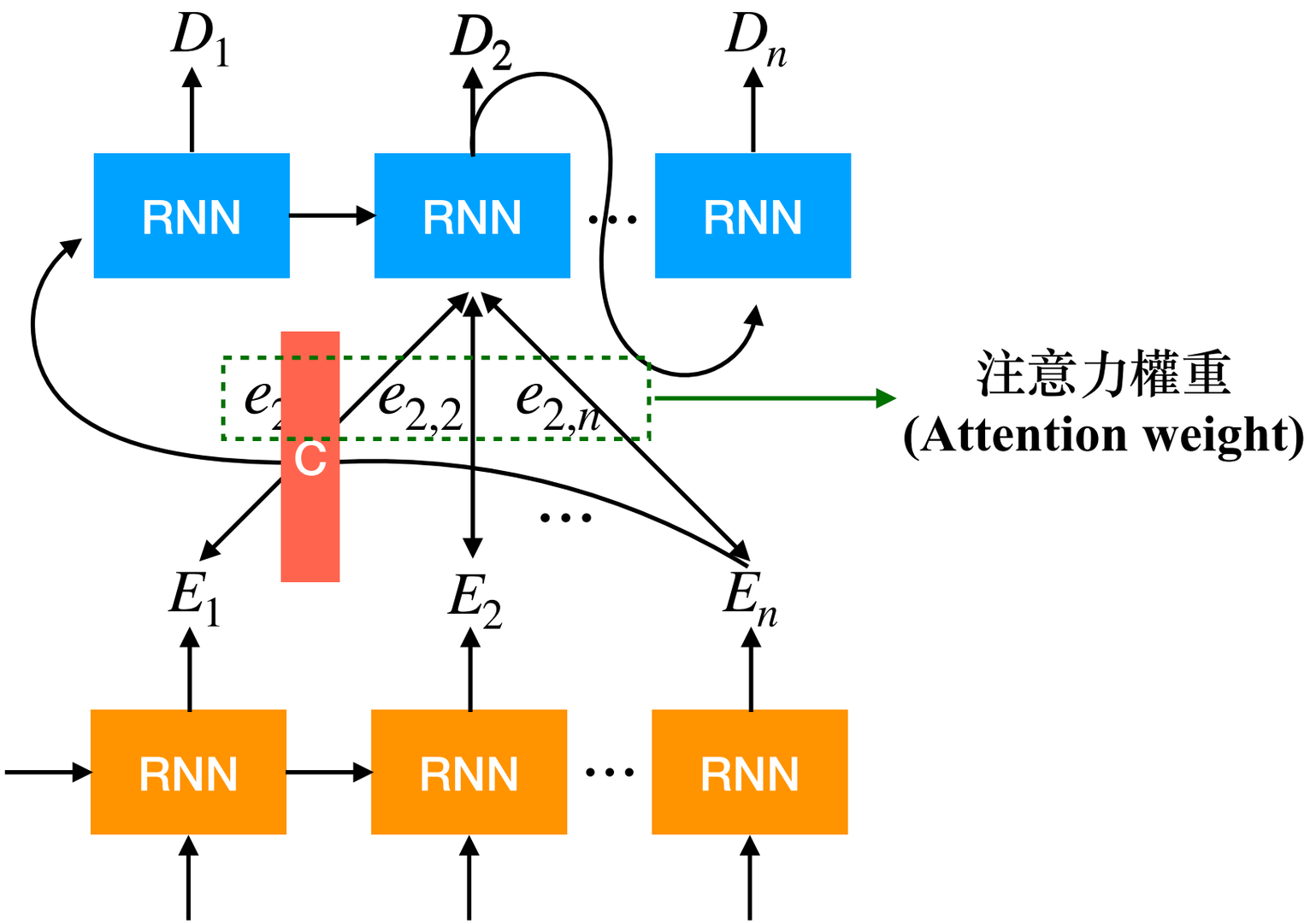
- **Fixed Size Context vector：**

對於編碼器來說，不管輸入序列有多長，最終都會被一個context vector(向量)來表示。

對於解碼器來說，只有第一個RNN會接收到編碼器傳遞過來的資訊(context vector)，之後會漸漸的遺忘。

Seq2seq with Attention

解碼器：



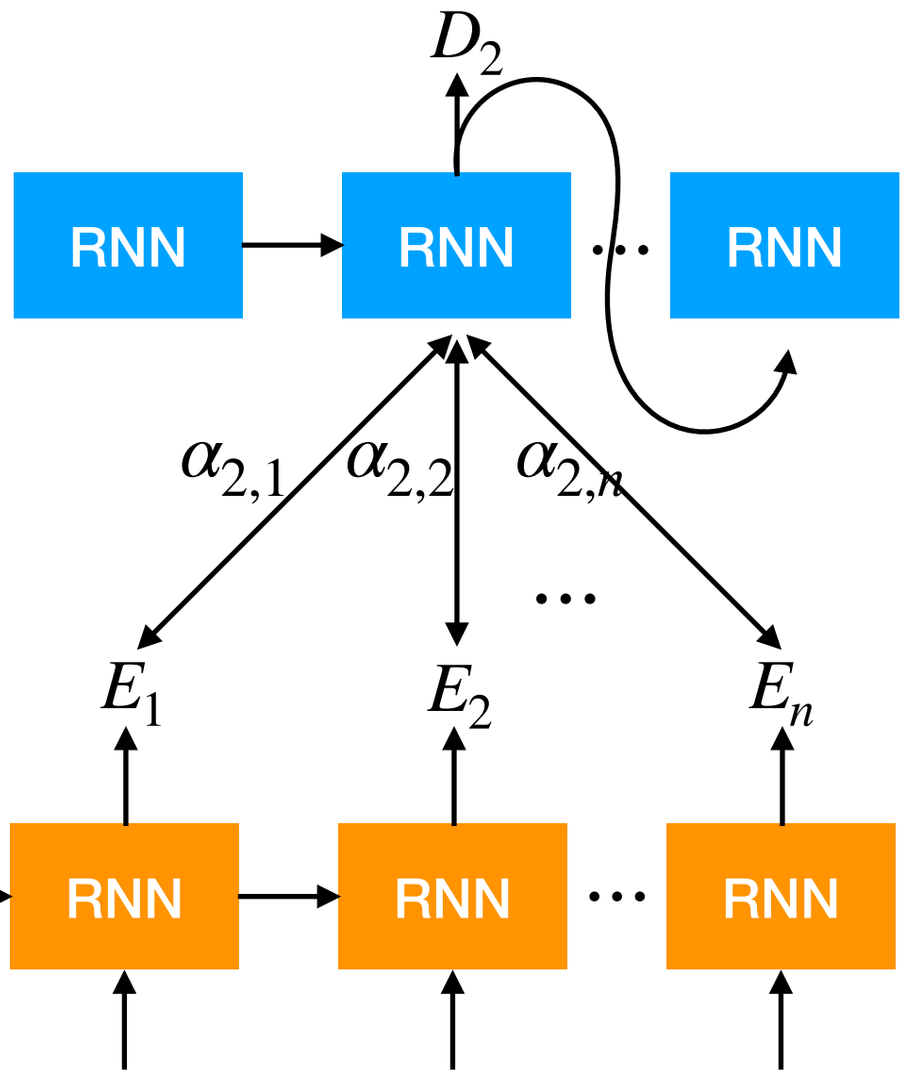
編碼器：

$$e_{i,j} = f(E_i, D_j) = \begin{cases} E_i D_j^T, & dot \\ E_i W D_j, & general \\ \tanh(w^T(E_i W_E + D_j W_D)), & feedforward \end{cases}$$

推薦閱讀：https://medium.com/ai-academy-taiwan/attention-mechanism-fad735db3c2c

Seq2seq with Attention

解碼器：



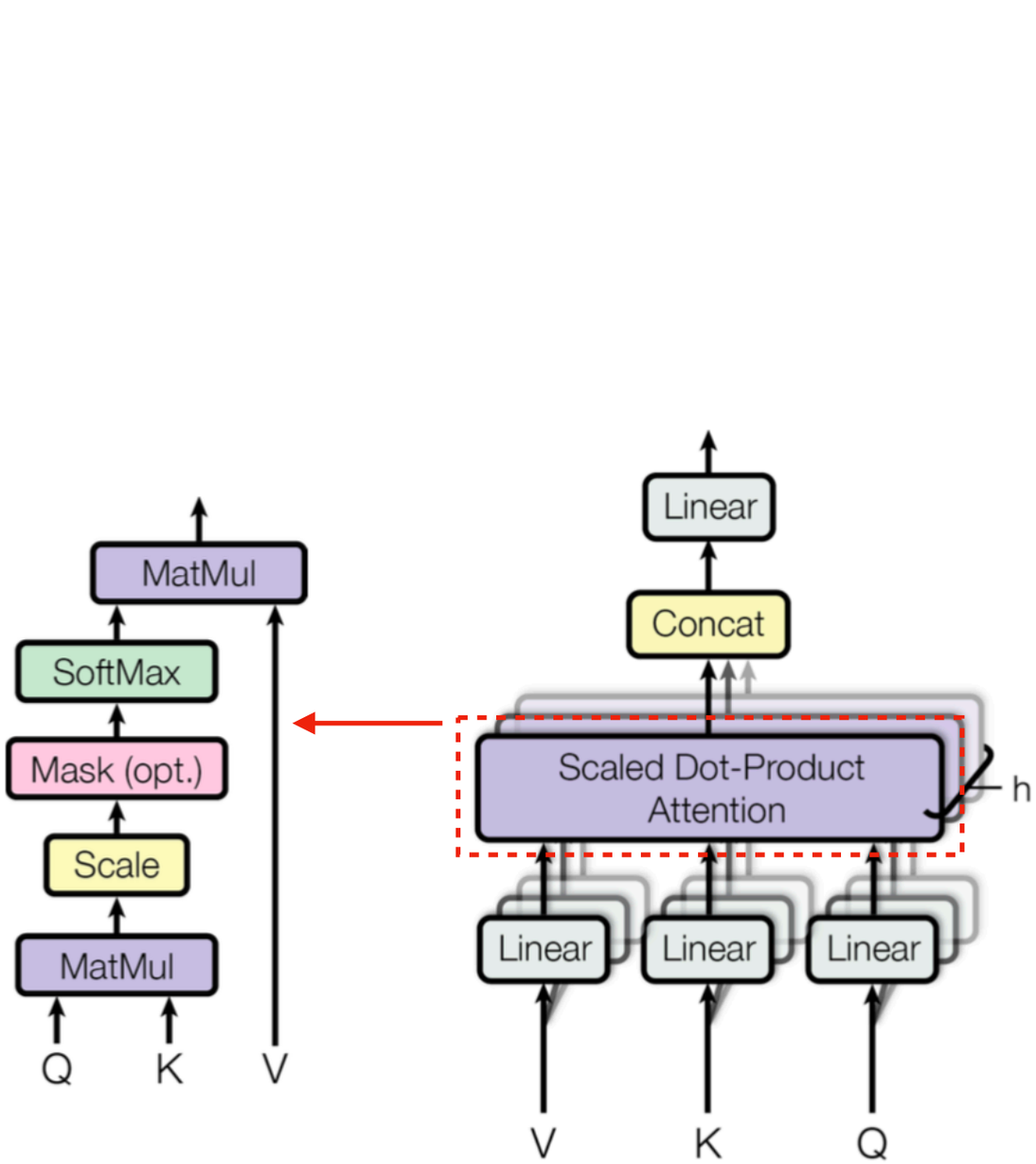
編碼器：

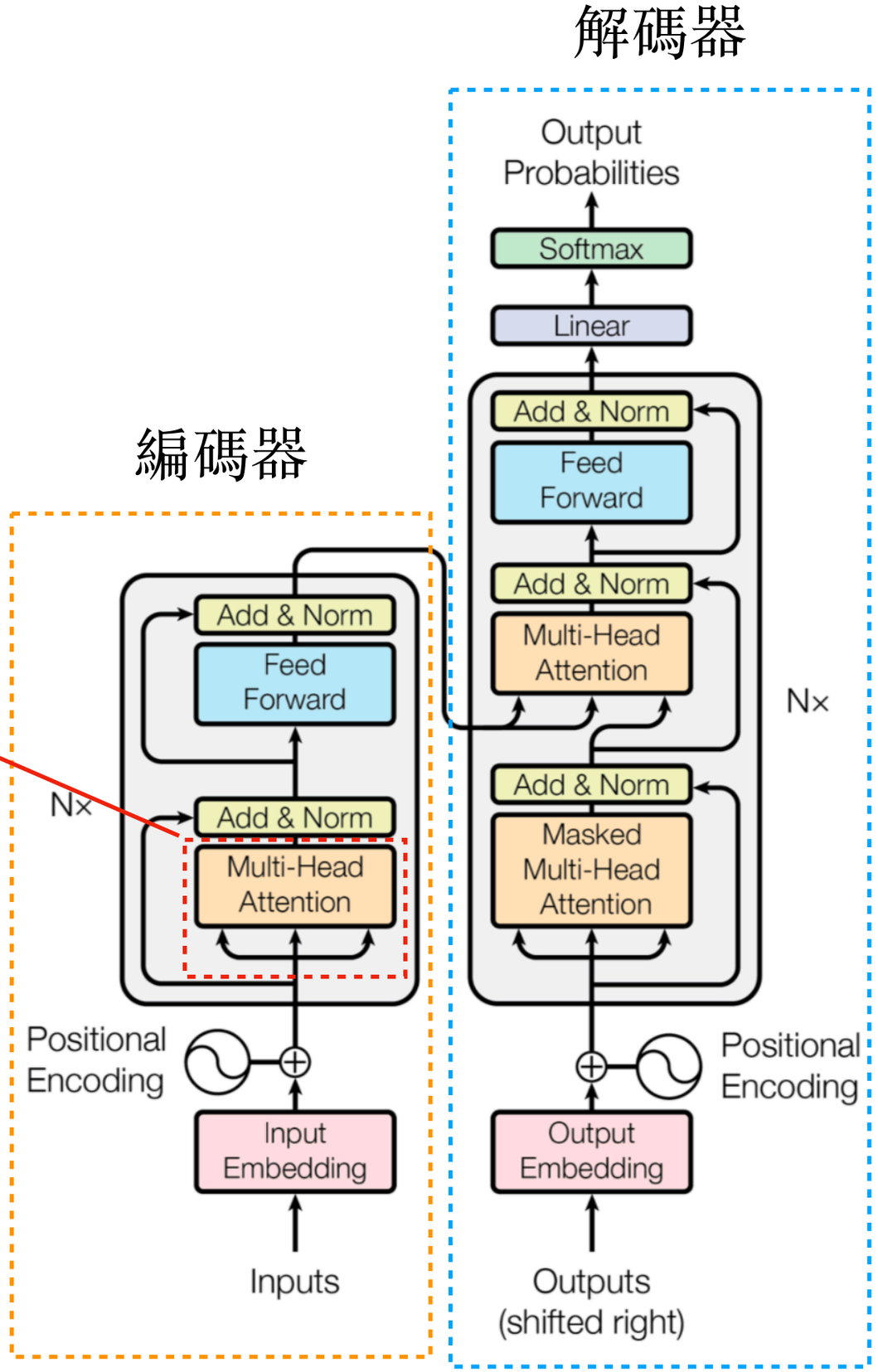
1. $\alpha_{ij} = softmax(e_{ij}|e) = \dfrac{\exp(e_{ij})}{\sum_{k=1}^{n}\exp(e_{ik})}$

2. $c_i = \sum_{j=1}^{n} a_{ij}E_i$

推薦閱讀：https://medium.com/ai-academy-taiwan/attention-mechanism-fad735db3c2c

Transformer

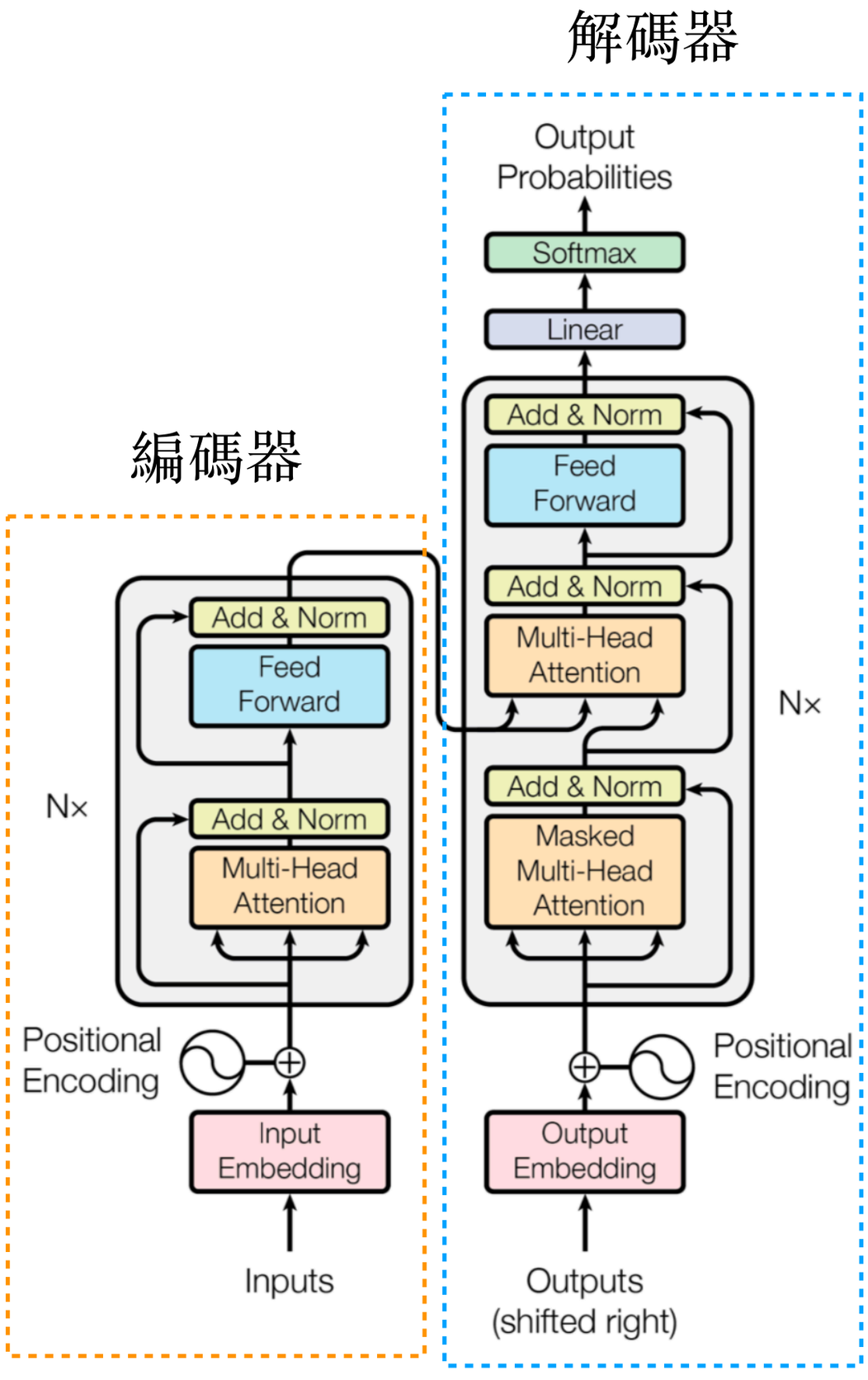




Transformer

Encoder 家族：

- BERT
- RoBERTa
- Transformer-XL
- ERNIE
- ELECTRA
- ALBERT
- DistilBERT
- XLM
- …



Decoder 家族：

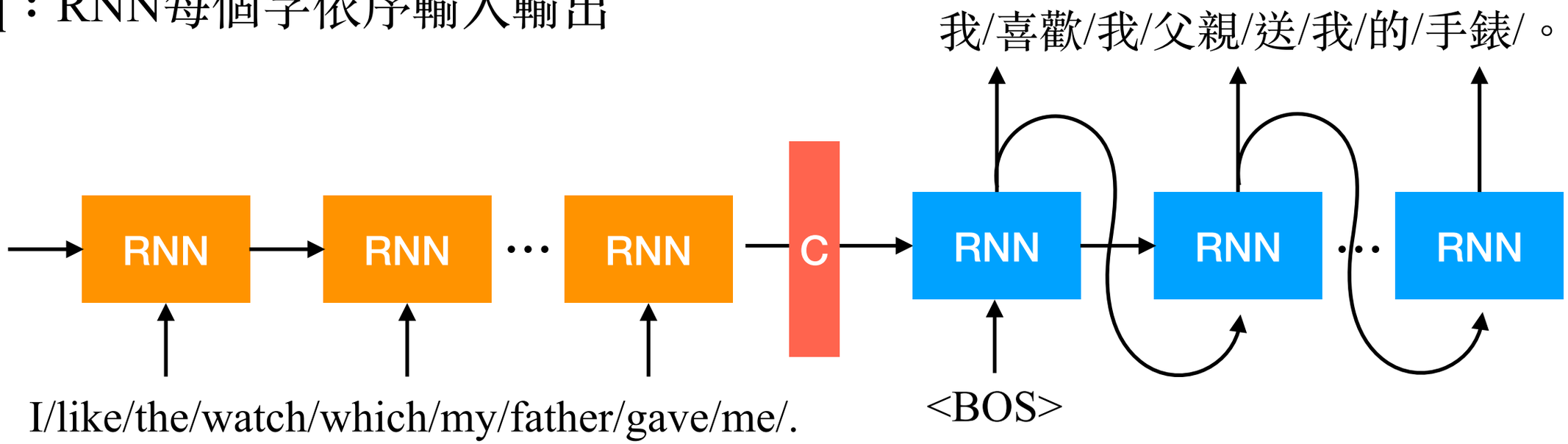
- GPT-1, 2, 3
- CTRL
- …

一起用：

- MASS
- mBART
- T5
- …

Comparison

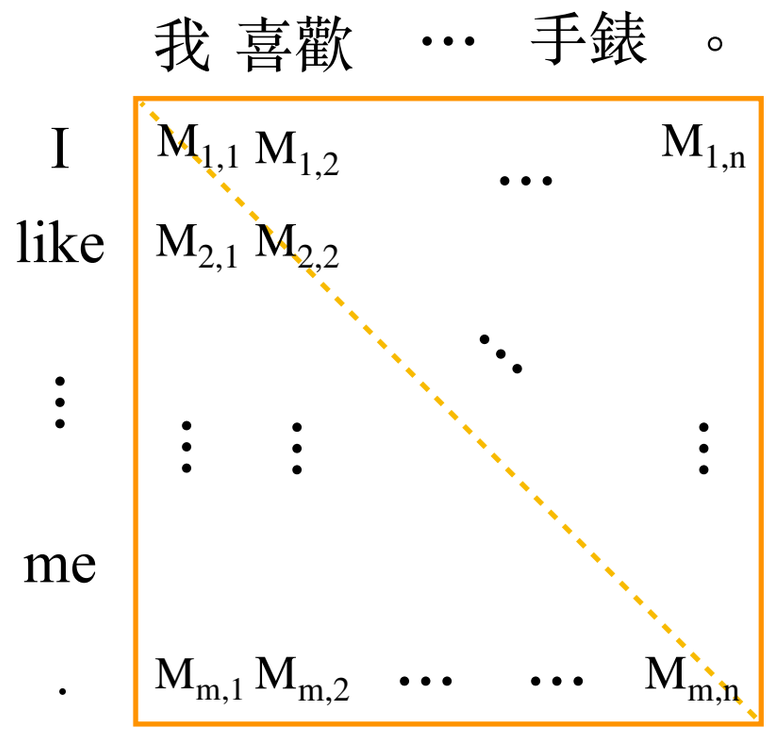
Seq2seq：RNN每個字依序輸入輸出

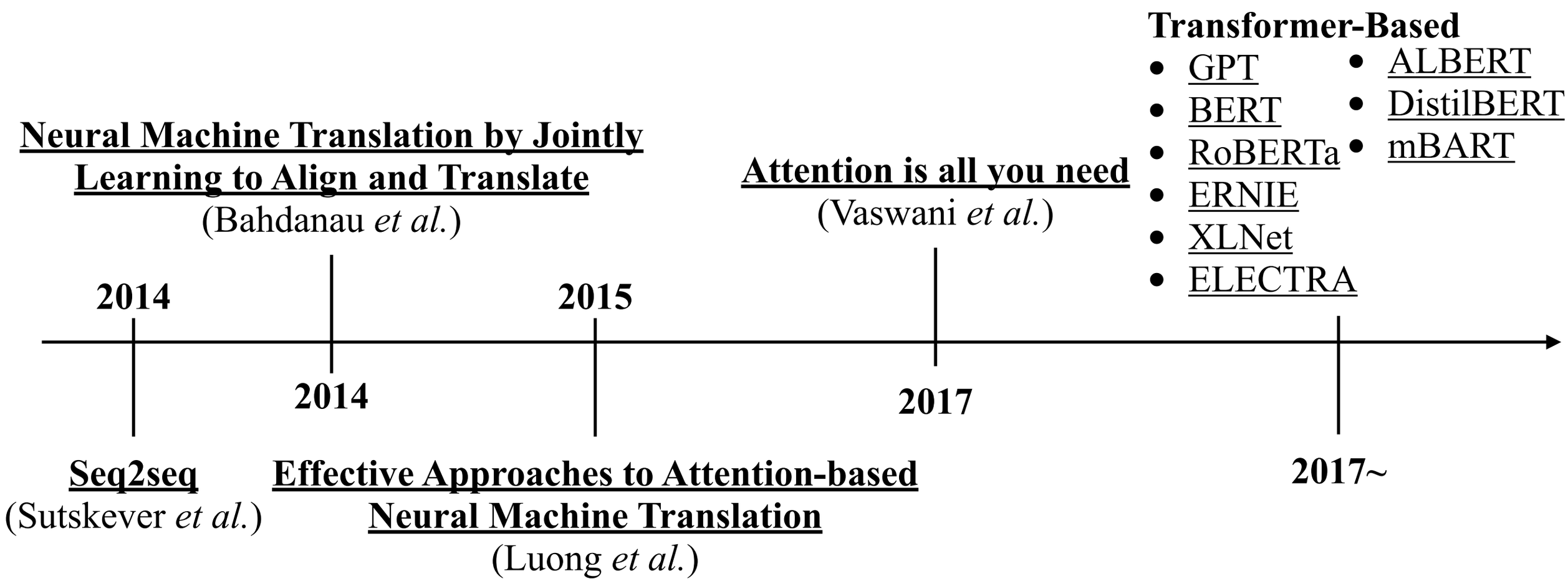


Transformer：

一次輸入所有的字進行Attention

(但沒有考慮到相鄰字之間的關係)





- Seq2seq：首次將類神經網路應用在翻譯任務(Neural Machine Translation)。
- Attention：設計網路結構時，考慮人類在翻譯時的思考邏輯(Soft-Attention)。
- Transformer：克服遞迴神經網絡(RNN)的缺點，大幅度提升翻譯水準。

Meta Learning

考驗(語言)模型的泛化性 (Generalization) 與流動性 (Fluidity)

Generalization：

- Training：

Sentence	Target
小明考試100分	Good
…	…

- Inference：

Sentence	Target
小王考試100分	?
…	…

Fluidity：

- Training：

Sentence	Target
1 + 1=	2
…	…

- Inference：

Sentence	Target
一加一等於	二
…	…

Zero, One, Few-Shot Learning

Large Corpus —Pre-train→ LM

- Encoder：Masked Language Model
- Decoder：AutoRegressive

- **Few-Shot Learning**：K examples + One prompt

Machine Translation

1. Artificial Intelligence → 人工智慧
2. Natural Language Processing → 自然語言處理
3. Machine Learning → 機器學習
k. …

Finetune → LM → Inference → • Computer Vision → ?

Zero, One, Few-Shot Learning

Large Corpus —Pre-train→ **LM**

- Encoder：Masked Language Model
- Decoder：AutoRegressive

- **One-Shot Learning**：One examples + One prompt

Machine Translation

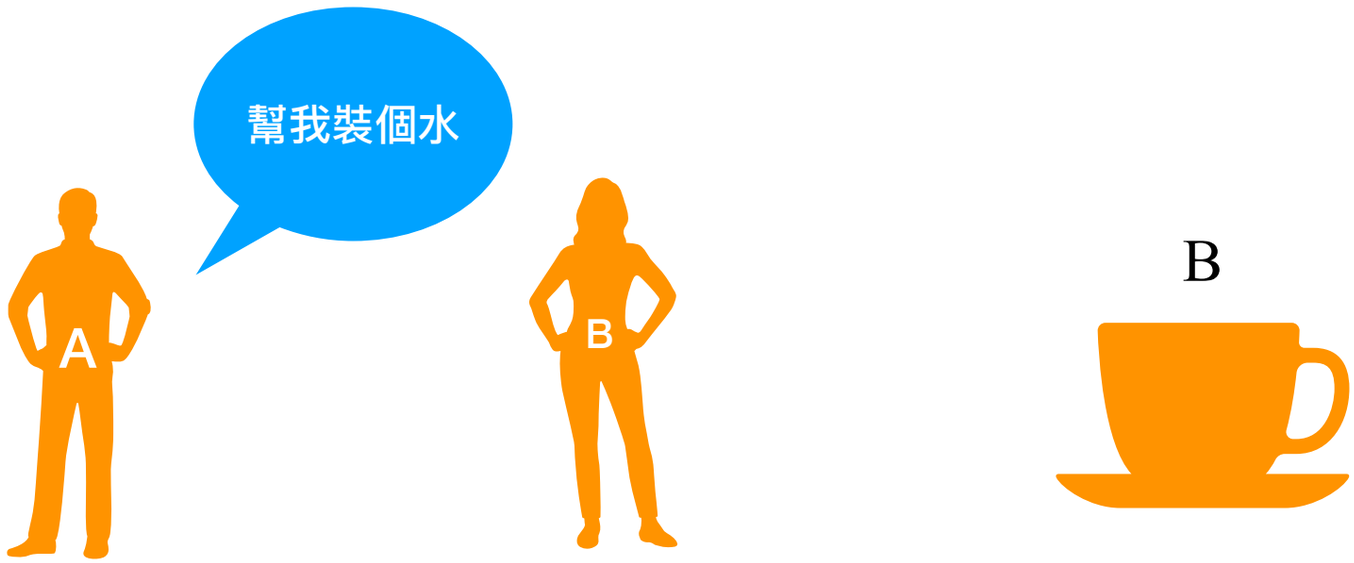
1. Artificial Intelligence → 人工智慧 —Finetune→ **LM** —Inference→ • Computer Vision → ?

Zero, One, Few-Shot Learning

Large Corpus → Pre-train → LM

- Encoder：Masked Language Model
- Decoder：AutoRegressive

- **Zero-Shot Learning**：One prompt



Machine Translation

LM → Inference → • Computer Vision → ?

Zero, One, Few-Shot Learning

- 資料量：

Few-Shot > One-Shot > Zero-Shot

- 下游任務表現：

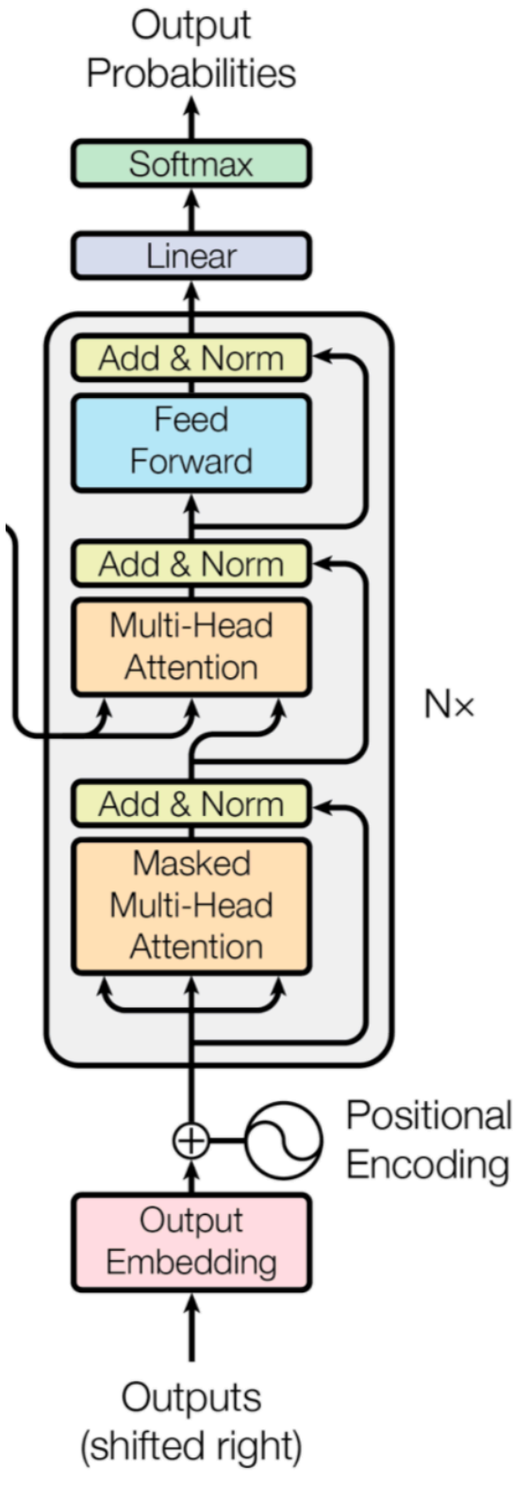
Few-Shot > One-Shot > Zero-Shot

- 考驗模型泛化能力：

Zero-Shot > One-Shot > Few-Shot

1. 哪一種比較符合人類的任務表達邏輯？
2. 資料量與泛化能力成正比？
3. 參數量與泛化能力成正比？
4. 大規模資料量 + 大規模參數量 = 泛化能力？

GPT-3
(Generative Pre-Training)



Training, Inference mode

T_1 : <BOS> → Model → 文

T_2 : <BOS>文 → Model → 瑋

⋮ ⋮

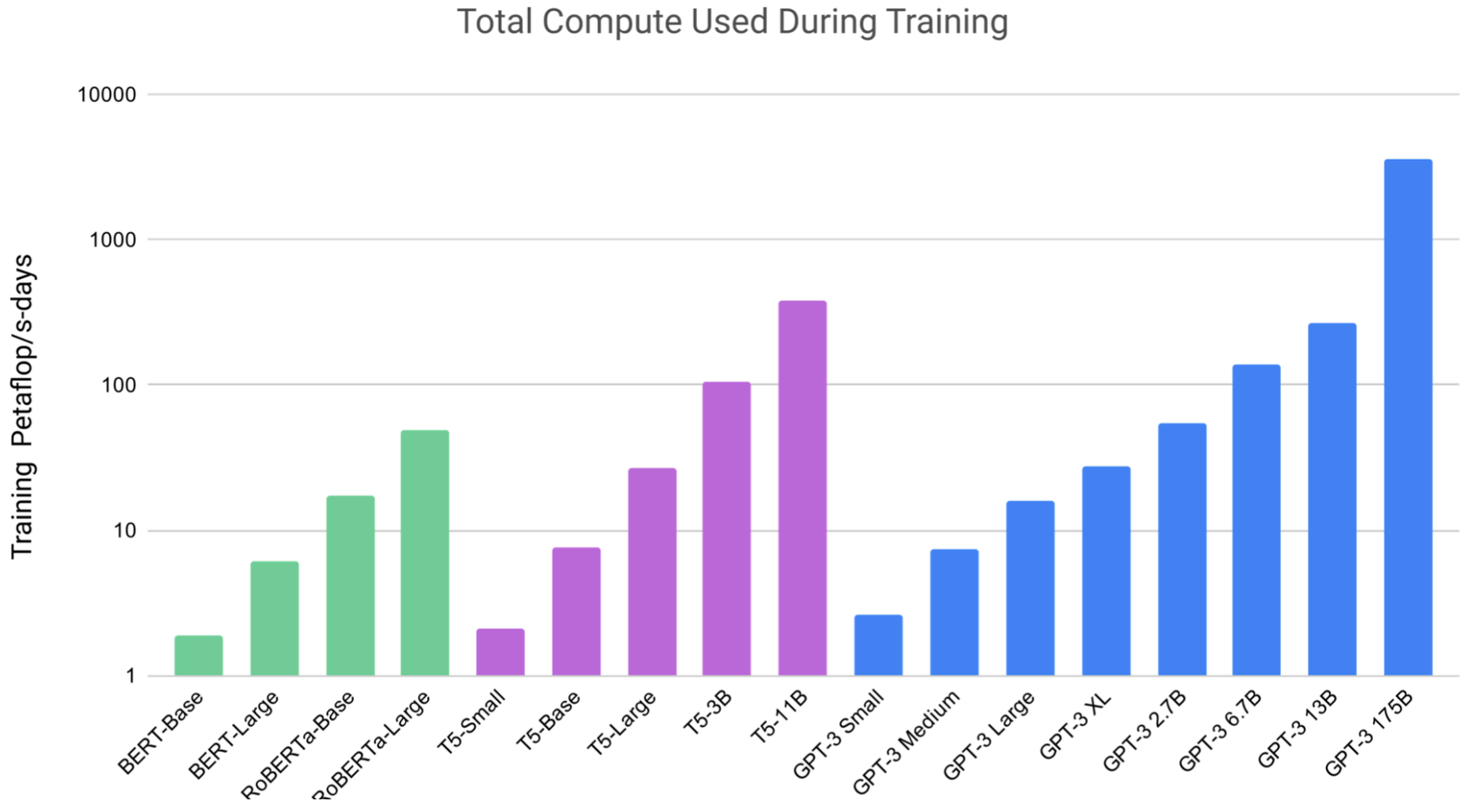
T_6 : <BOS>文瑋助教真 → Model → 帥

T_7 :<BOS>文瑋助教真帥 → Model → <EOS>

GPT-3

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

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



GPT-3

The three settings we explore for in-context learning

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

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

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

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

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

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

Traditional fine-tuning (not used for GPT-3)

Fine-tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.

```
1  sea otter => loutre de mer          ← example #1
```

↓

gradient update

↓

```
1  peppermint => menthe poivrée        ← example #2
```

↓

gradient update

↓ ↓ ↓

```
1  plush giraffe => girafe peluche     ← example #N
```

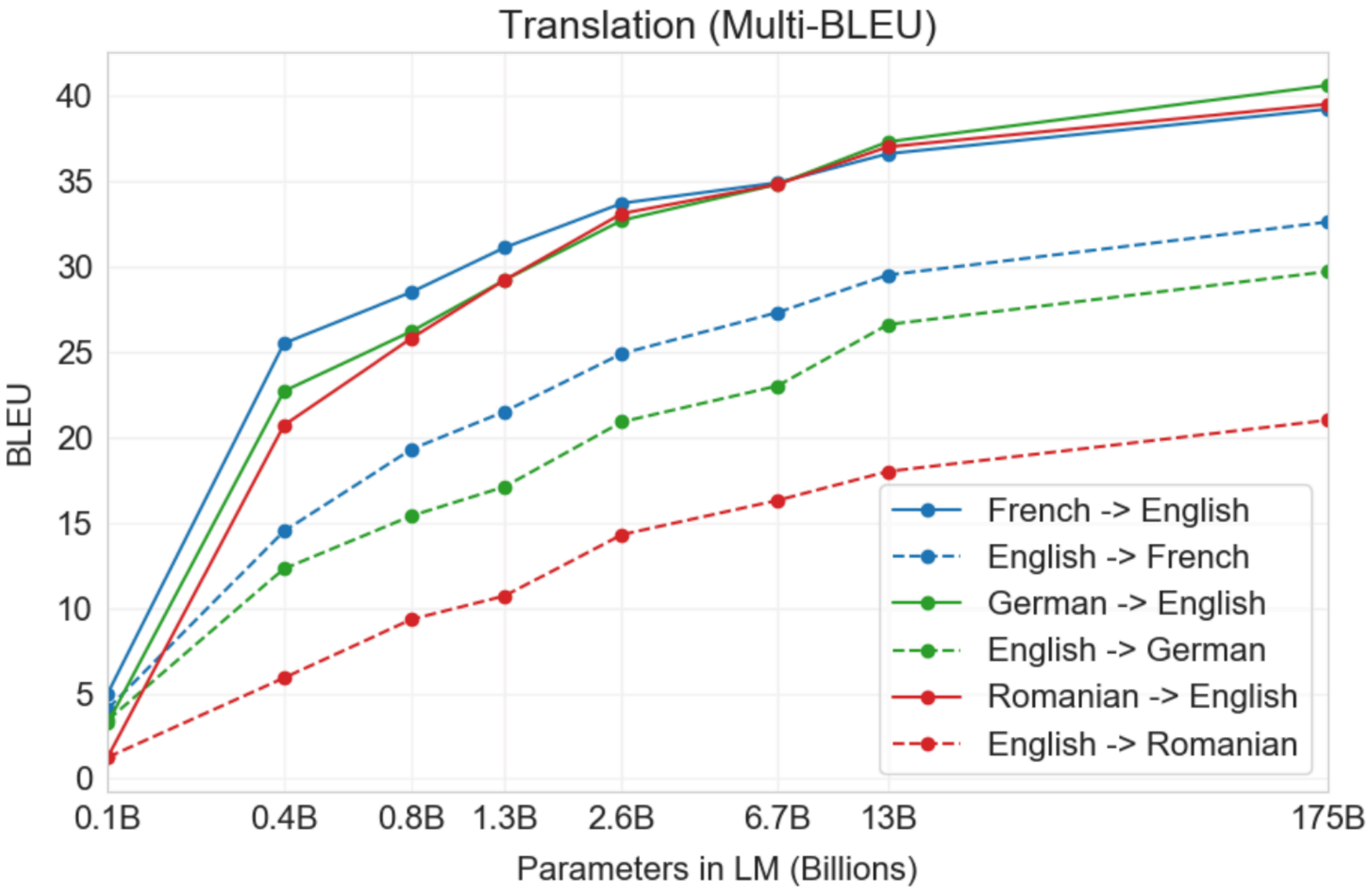
gradient update

```
1  cheese =>                           ← prompt
```

GPT-3

Setting	En→Fr	Fr→En	En→De	De→En	En→Ro	Ro→En
SOTA (Supervised)	**45.6**[a]	35.0 [b]	**41.2**[c]	40.2[d]	**38.5**[e]	**39.9**[e]
XLM [LC19]	33.4	33.3	26.4	34.3	33.3	31.8
MASS [STQ+19]	37.5	34.9	28.3	35.2	35.2	33.1
mBART [LGG+20]	-	-	29.8	34.0	35.0	30.5
GPT-3 Zero-Shot	25.2	21.2	24.6	27.2	14.1	19.9
GPT-3 One-Shot	28.3	33.7	26.2	30.4	20.6	38.6
GPT-3 Few-Shot	32.6	39.2	29.7	40.6	21.0	39.5

- XLM (Cross-lingual Language Model)：Pre-train and fine-tune on multi-language.
- MASS (Masked Sequence to Sequence)：Fine-tune on downstream task.
- mBART：Pre-train and fine-tune on multi-language.
- GPT-3：Without fine-tune.


Translation (Multi-BLEU)
BLEU
40
35
30
25
20
15
10
5
0
0.1B
0.4B
0.8B
1.3B
2.6B
6.7B
13B
175B
Parameters in LM (Billions)
French -> English
English -> French
German -> English
English -> German
Romanian -> English
English -> Romanian

GPT-3

Setting	PIQA	ARC (Easy)	ARC (Challenge)	OpenBookQA
Fine-tuned SOTA	79.4	**92.0**[KKS+20]	**78.5**[KKS+20]	**87.2**[KKS+20]
GPT-3 Zero-Shot	**80.5***	68.8	51.4	57.6
GPT-3 One-Shot	**80.5***	71.2	53.2	58.8
GPT-3 Few-Shot	**82.8***	70.1	51.5	65.4

- PIQA : PhysicalQA

Context →	How to apply sealant to wood.
Correct Answer →	Using a brush, brush on sealant onto wood until it is fully saturated with the sealant.
Incorrect Answer →	Using a brush, drip on sealant onto wood until it is fully saturated with the sealant.

Figure G.4: Formatted dataset example for PIQA

GPT-3

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

RACE-m

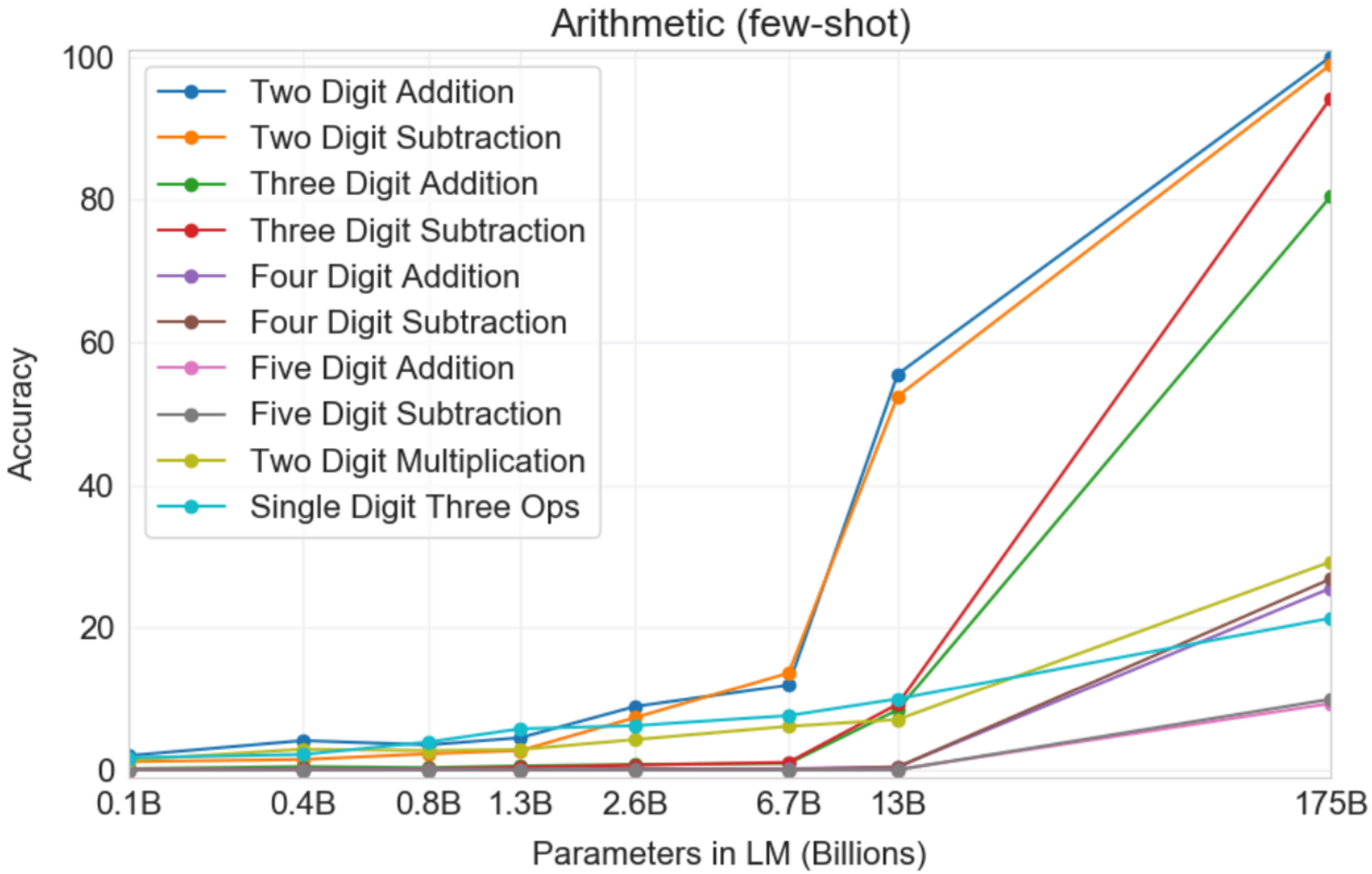
Few-Shot

One-Shot

Zero-Shot

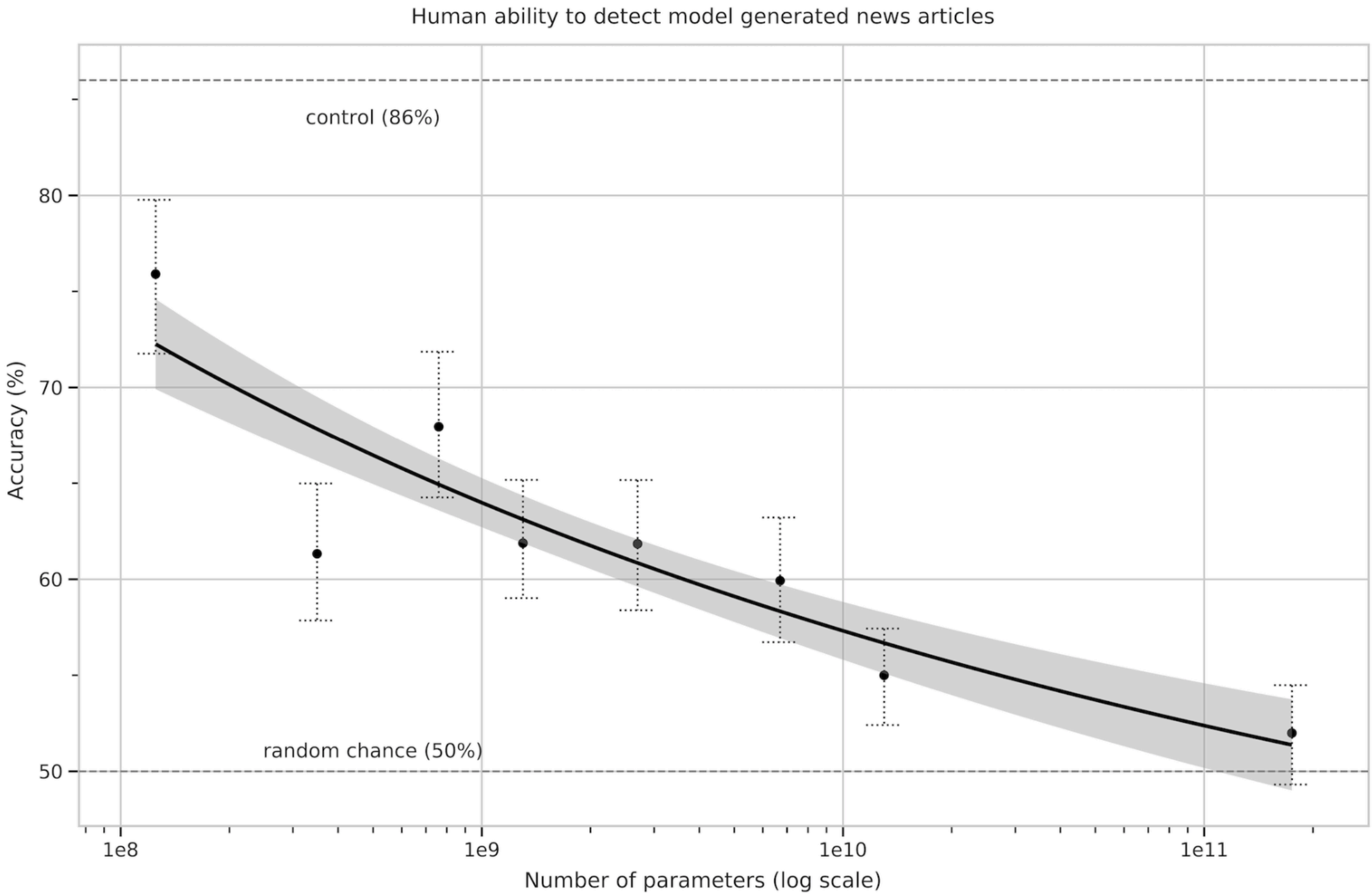
```
Context →  Article:
Mrs.  Smith is an unusual teacher.  Once she told each student to bring
along a few potatoes in plastic bag.  On each potato the students had to
write a name of a person that they hated And the next day, every child
brought some potatoes.  Some had two potatoes;some three;some up to five.
Mrs.  Smith then told the children to carry the bags everywhere they went,
even to the toilet, for two weeks.  As day after day passed, the children
started to complain about the awful smell of the rotten potatoes.
Those children who brought five potatoes began to feel the weight trouble
of the bags.  After two weeks, the children were happy to hear that the
game was finally ended.  Mrs.  Smith asked,"How did you feel while carrying
the potatoes for two weeks?" The children started complaining about the
trouble loudly.
Then Mrs.  Smith told them why she asked them to play the game.  She
said,"This is exactly the situation when you carry your hatred for somebody
inside your heart.  The terrible smell of the hatred will pollute your
heart and you will carry something unnecessary with you all the time.  If
you cannot stand the smell of the rotten potatoes for just two weeks, can
you imagine how heavy it would be to have the hatred in your heart for your
lifetime?  So throw away any hatred from your heart, and you'll be really
happy."

Q: Which of the following is True according to the passage?

A: If a kid hated four people,he or she had to carry four potatoes.

Q: We can learn from the passage that we should _ .

A: throw away the hatred inside

Q: The children complained about _ besides the weight trouble.

A: the smell

Q: Mrs.Smith asked her students to write _ on the potatoes.

A:

Correct Answer →   names
Incorrect Answer →  numbers
Incorrect Answer →  time
Incorrect Answer →  places
```

GPT-3



Two Digit Addition：Q:What is 48 plus 76? A:

GPT-3


Human ability to detect model generated news articles
control (86%)
random chance (50%)
Accuracy (%)
80
70
60
50
1e8
1e9
1e10
1e11
Number of parameters (log scale)

GPT-3

Title：Star’s Tux Promise Draws Megyn Kelly’s Sarcasm

Subtitle：Joaquin Phoenix pledged to not change for each awards event

Article：A year ago, Joaquin Phoenix made headlines when he appeared on the red carpet at the Golden Globes wearing a tuxedo with a paper bag over his head that read, “I am a shape-shifter. I can only change myself.” It was a promise to not change to fit into the Hollywood mold: “I think it’s a really special thing to say, ‘This is what’s inside of me, I’m proud of it, and I’m not going to be ashamed because of the way that someone else thinks I should be.” Now, it’s the Oscars, and Phoenix is at it again. But this time, his publicist is saying he’ll be wearing a tux no matter what.

Title：Star’s Tux Promise Draws Megyn Kelly’s Sarcasm

Subtitle：Joaquin Phoenix pledged to not change for each awards event

Article：PETA praised Joaquin Phoenix, actor and climate change activist, for announcing that he'll wear the same tuxedo to every Hollywood awards event this season to reduce waste. Stella McCartney, who designed his classic black tux, also was impressed. Former Fox News and NBC host Megyn Kelly was not, People reports. "How out of touch are these ppl?? Oh the sacrifice!" Kelly tweeted after a woman posted that her husband wears the same firefighter ensemble every day.