

大型語言模型的成長史

講者：李宏毅

大型語言模型的成長史



Pre-train



(Instruction)
Fine-tuning



Reinforcement Learning from
Human Feedback (RLHF)

背景知識：文字接龍

原本的目標

臺灣最高的山是哪座？



語言模型



玉山

拆解成一連串文字接龍

臺灣最高的山是哪座？



語言模型



玉

token

臺灣最高的山是哪座？玉



語言模型



山

臺灣最高的山是哪座？玉山



語言模型



[END]

背景知識：機器學習/深度學習

模型

測試, testing (推論, inference)

上萬個參數

有一個函式 貓 or 狗 = $f(\text{)}$ = $\overbrace{\dots a \dots b \dots c \dots d \dots e \dots f \dots g \dots}^{\text{上萬個參數}}$
 $a = 1.3, b = -7.2, c = 0.4, \dots$



輸出： 貓



輸出： 狗



輸出： 貓



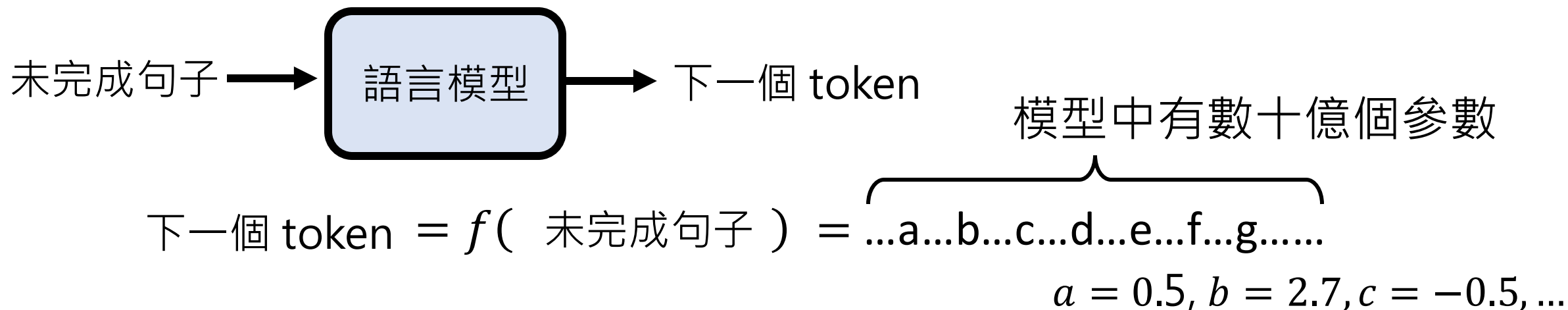
輸出： 狗

訓練資料

機器學習可以把上萬個參數找出來 訓練, training (學習, learning)

$a = 1.3, b = -7.2, c = 0.4, \dots$

背景知識：機器學習/深度學習套用到文字接龍上



訓練資料

輸入：人工智	輸出：慧
輸入：不要忘了今天來開	輸出：會
輸入：床前明月	輸出：光
⋮	⋮

機器學習可以把數十億個參數找出來 $a = 0.5, b = 2.7, c = -0.5, \dots$

在語言模型成長的故事中



Pre-train



(Instruction)
Fine-tuning



Reinforcement Learning from
Human Feedback (RLHF)

所有階段的學習方法都是一樣的，只是訓練資料不同

機器學習找參數的挑戰



輸出： 貓



輸出： 狗



輸出： 貓

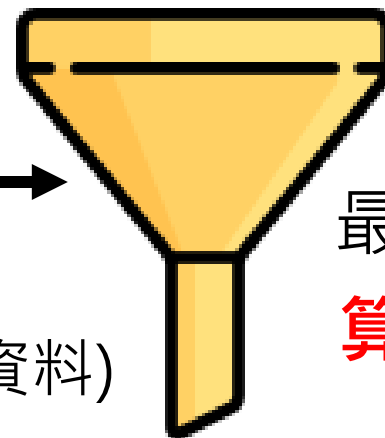


輸出： 狗

訓練資料

超參數
(hyperparameter)

設定



最佳化 (Optimization)

算力

最佳化可能會失敗 (找到的參數沒有符合訓練資料)

參數 $a = 1.3, b = -7.2, c = 0.4, \dots$

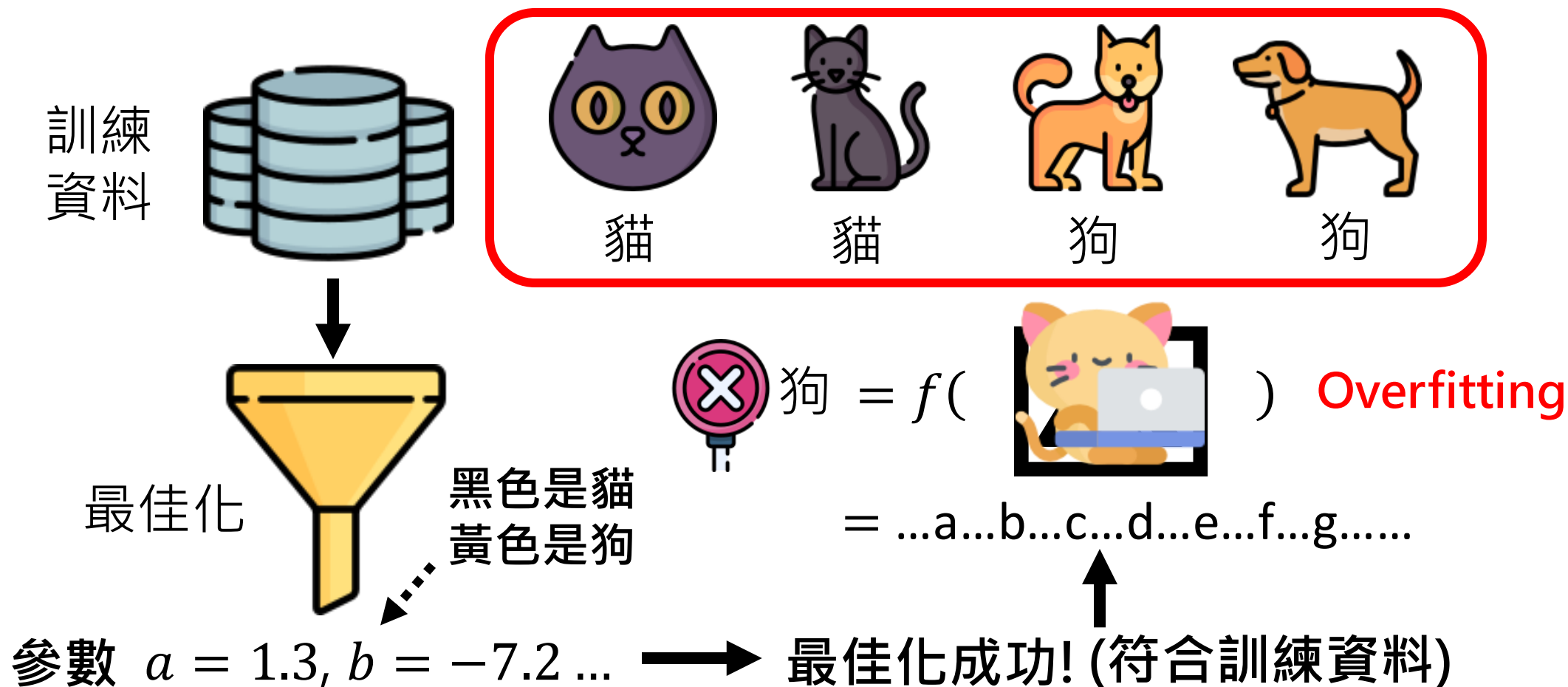
機器學習找參數的挑戰

- 最佳化失敗怎麼辦？換一組超參數再上一次！



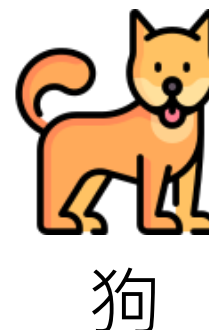
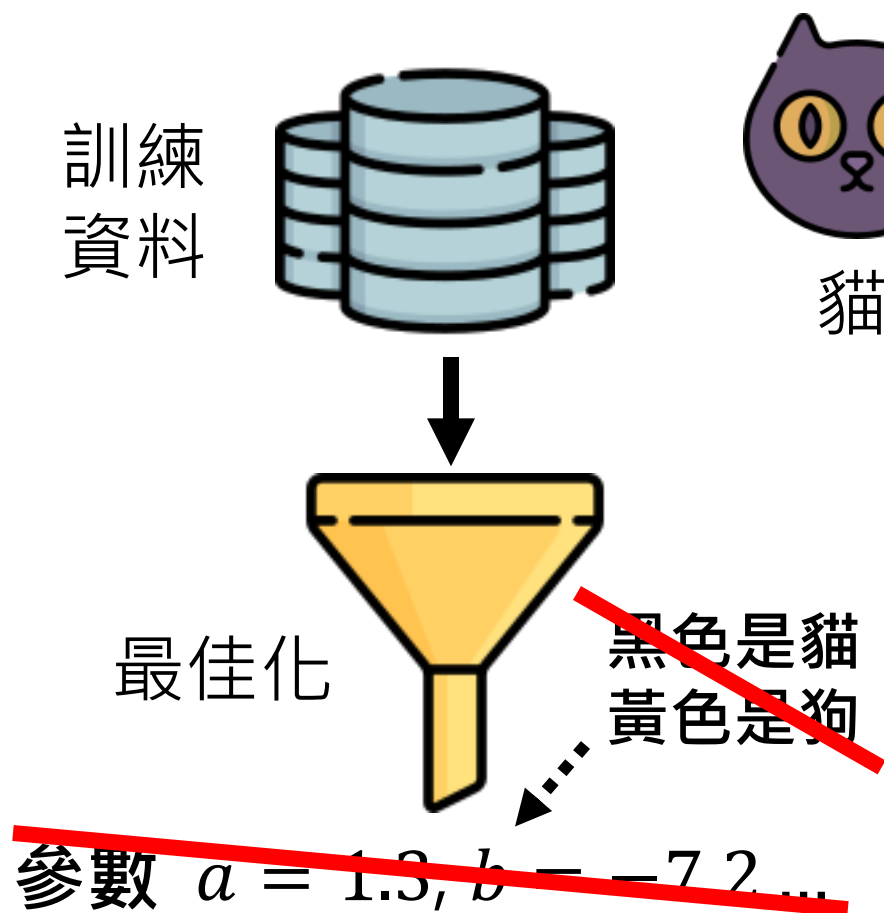
機器學習找參數的挑戰

最佳化只管有沒有「符合」訓練資料，不管有沒有道理



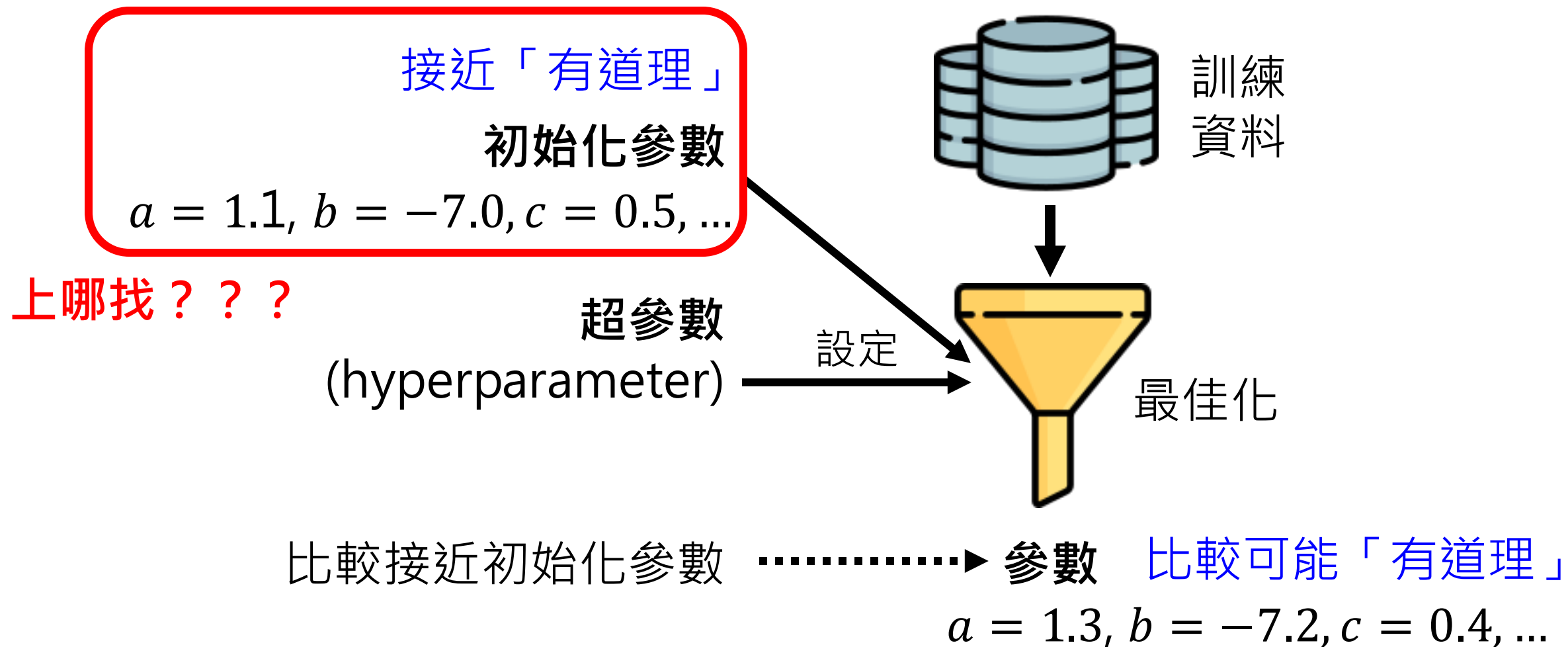
機器學習找參數的挑戰

最佳化只管有沒有「符合」訓練資料，不管有沒有道理



增加訓練資料的多樣性

如何讓機器找到比較「有道理」的參數



隨機初始化參數叫 “train from scratch”

大型語言模型的成長史



Pre-train



(Instruction)
Fine-tuning



Reinforcement Learning from
Human Feedback (RLHF)

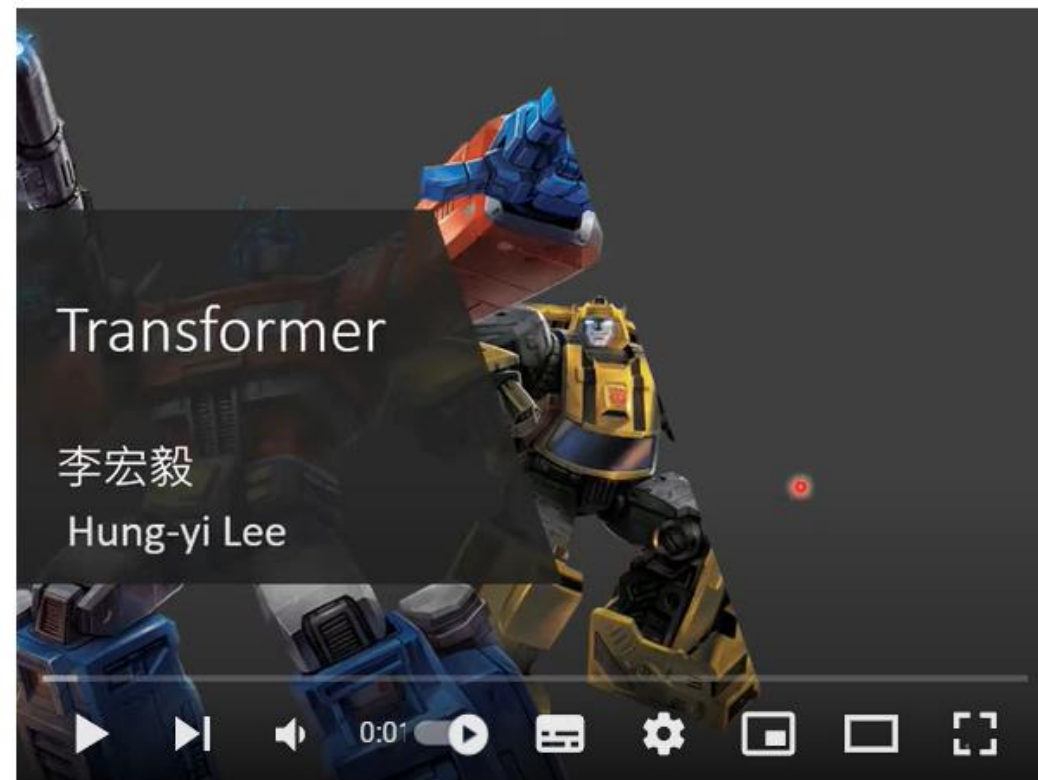
Pre-train 對 ChatGPT 非常重要 ...

ChatGPT

G: Generative 生成

P: Pre-trained 預訓練

T: Transformer



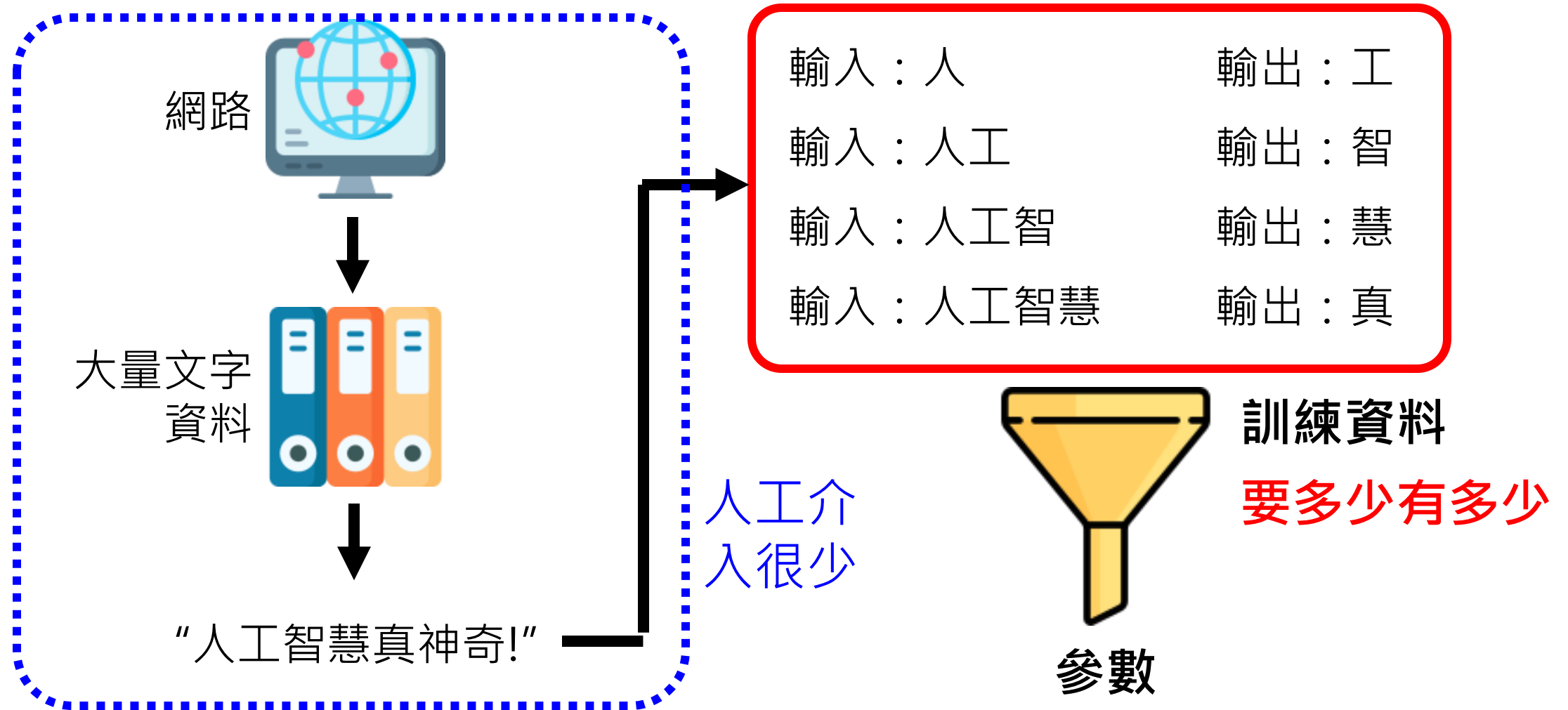
【機器學習2021】(中文版)

【機器學習2021】Transformer (上)

<https://youtu.be/n9TlOhRjYoc>

從任何文字資料學習

Self-supervised Learning (自督導式學習)



需要多少文字才夠？

When Do You Need Billions of Words of Pretraining Data?

<https://arxiv.org/abs/2011.04946>

語言知識

這個人突然就

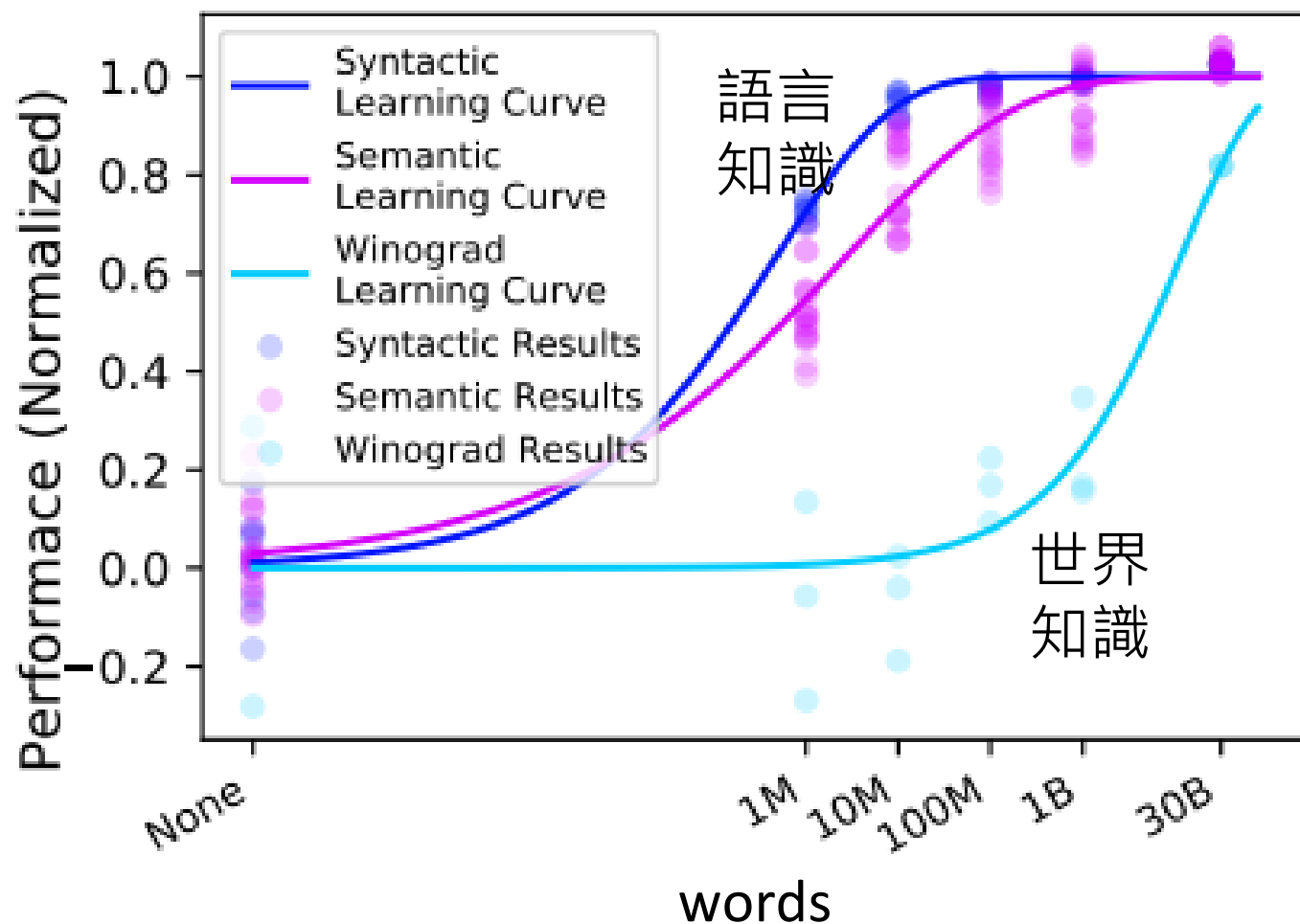
- 跑 ☒
- 飛 ☒
- 的 ☒

世界知識

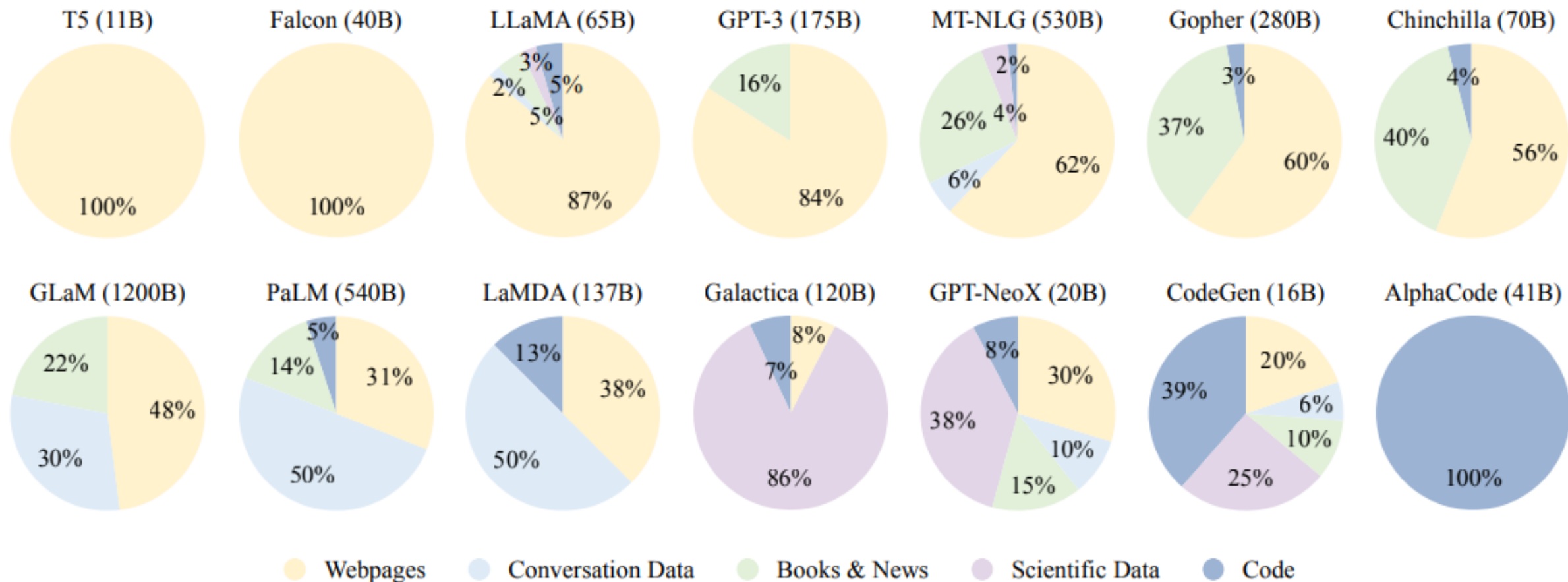
水的沸點是攝氏

- 一百度 ☒ ☒
- 五十度 ☒ ☒

在低壓下



預訓練資料來源



資料清理

- GPT-3/The Pile/PaLM 使用「資料品質」分類器
- 高品質的文句在資料訓練會被多次重覆



The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work

Millions of articles from The New York Times were used to train chatbots that now compete with it, the lawsuit said.

<https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html>

OpenAI and journalism

We support journalism, partner with news organizations, and believe The New York Times lawsuit is without merit.

<https://openai.com/blog/openai-and-journalism>

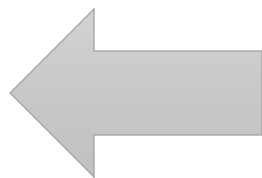
在 ChatGPT 之前的 GPT 系列

Model size:

**GPT-1
(2018)**



117M



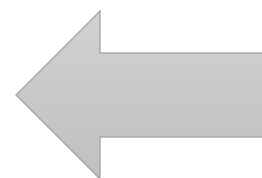
函式的參數量(複雜程度)

人工智慧的天資

Data size:

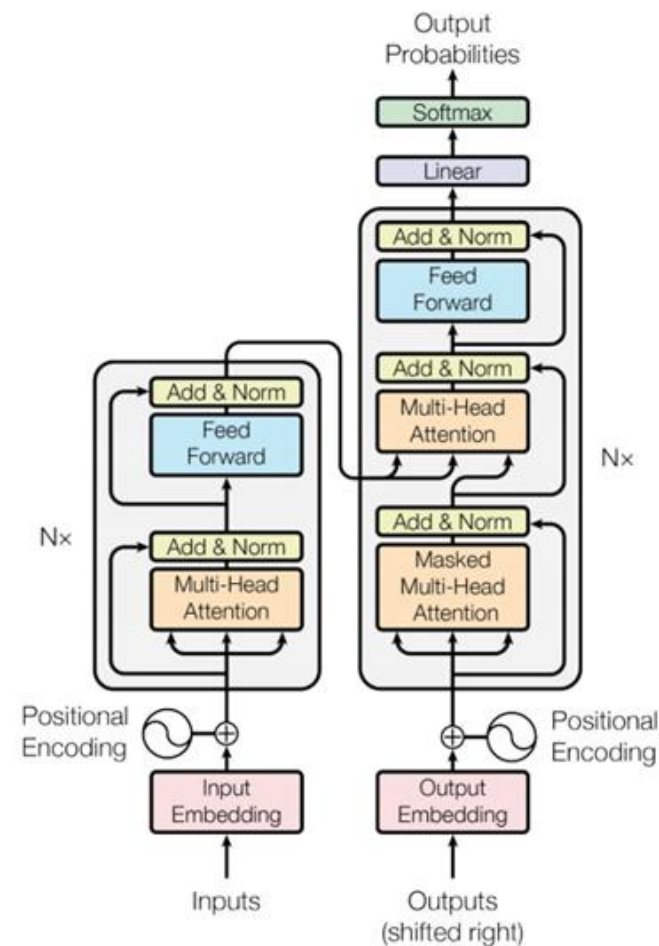


1GB



拿來學文字接龍的資料量

後天的努力



在 ChatGPT 之前的 GPT 系列

Model size:

GPT-1
(2018) 
117M

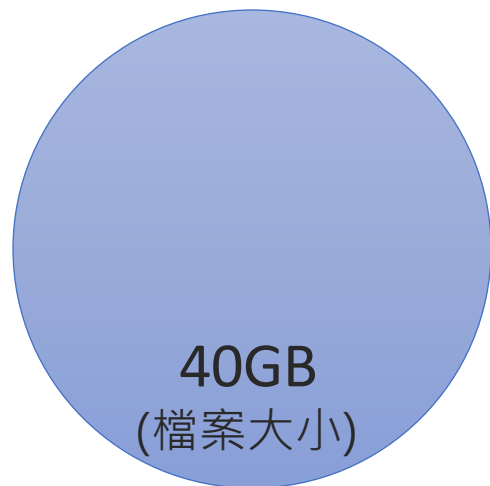
GPT-2
(2019)



1542M

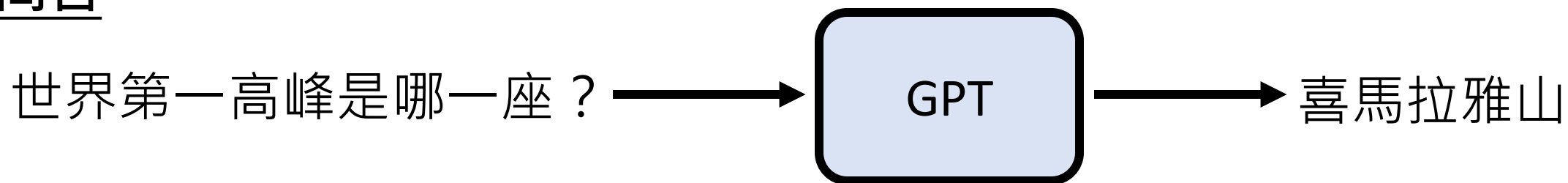
Data size:


7000
books

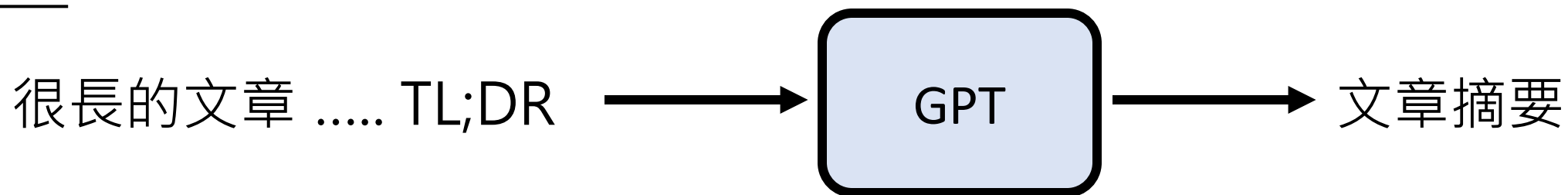

40GB
(檔案大小)

GPT-2 也是可以回答問題的!

問答



摘要

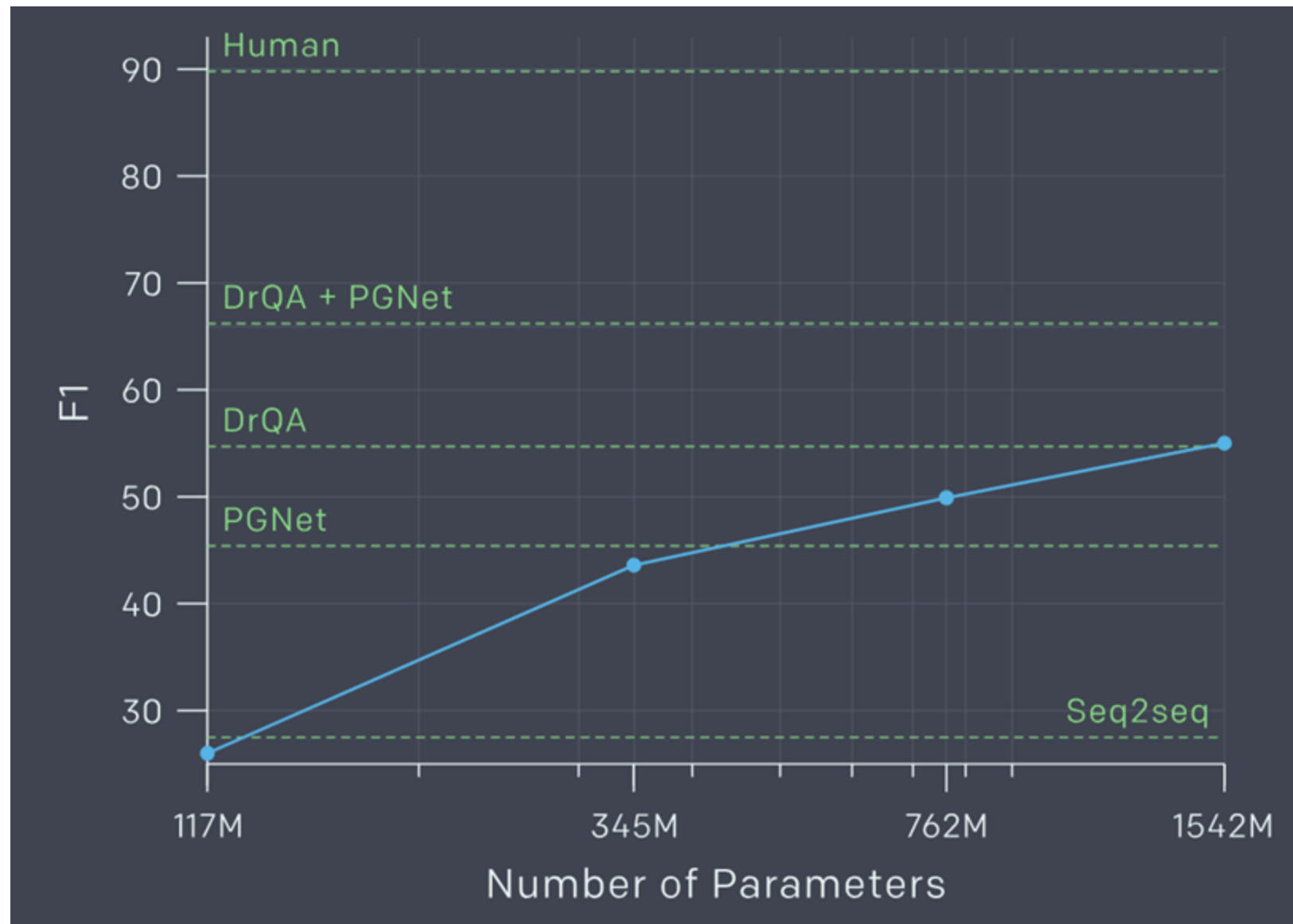


GPT-2

<https://openai.com/blog/better-language-models/>

問答上表現
如何？

CoQA



在 ChatGPT 之前的 GPT 系列

Model size:

GPT-1
(2018)



117M

GPT-2
(2019)



1542M

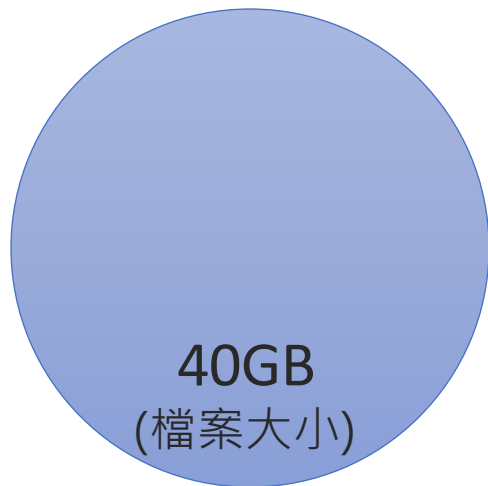
GPT-3
(2020)

175B

Data size:



7000
books

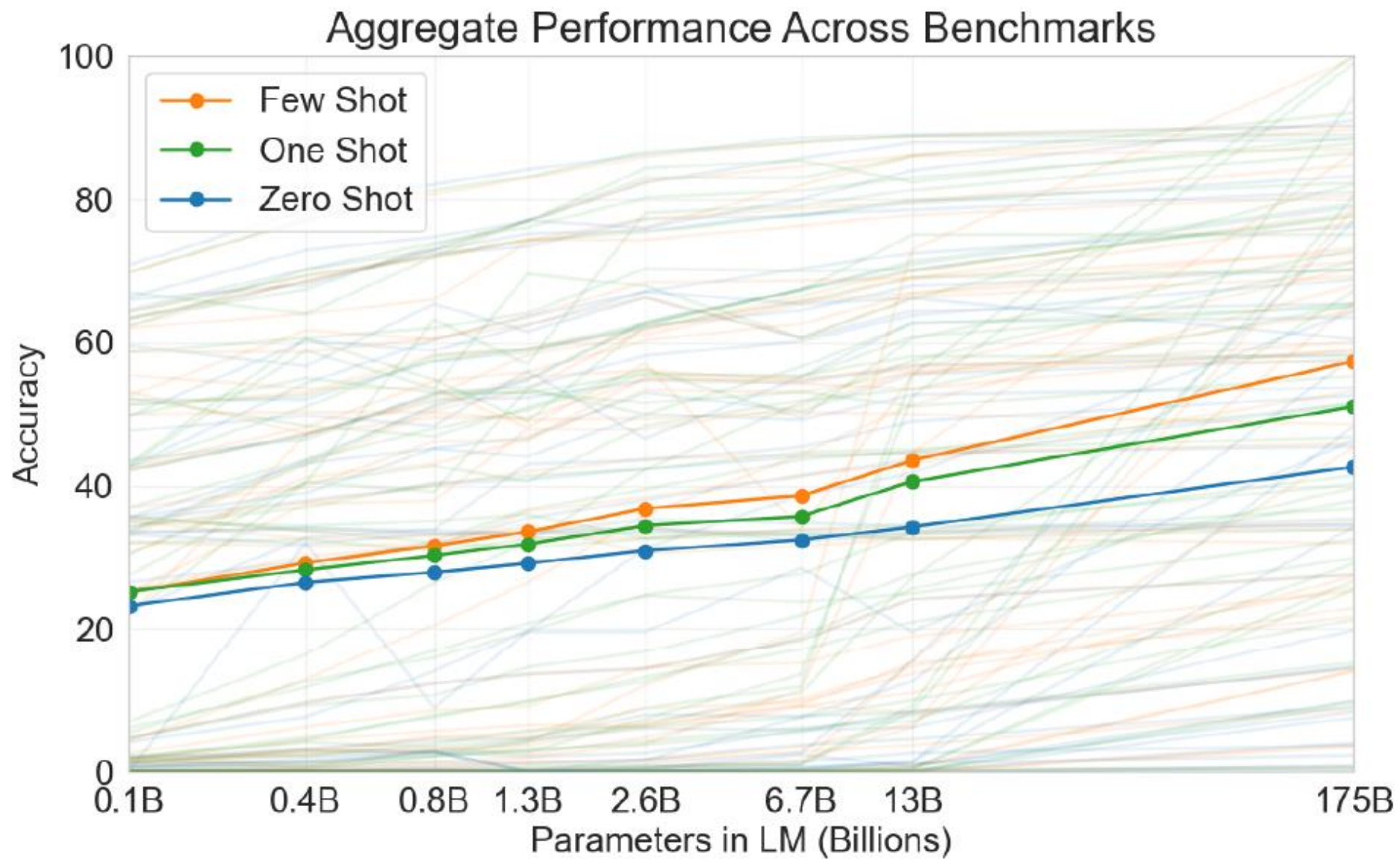


580GB (檔案大小)

300B tokens

閱讀哈利波特全集 30 萬遍

GPT-3



<https://arxiv.org/abs/2005.14165>

從任意大量文字資料學文字接龍，真能回答問題？

What is the purpose of the list C in the code below?

```
def binomial_coefficient(n, r):  
    C = [0 for i in range(r + 1)];  
    C[0] = 1;  
    for i in range(1, n + 1):  
        j = min(i, r);  
        while j > 0:  
            C[j] += C[j - 1];  
            j -= 1;  
    return C[r]
```

GPT-3

- A. to store the value of C[0]
- B. to store the value of C[1]
- C. to store the value of C[i]
- D. to store the value of C[i - 1]

還能不能更大？

Switch Transformer

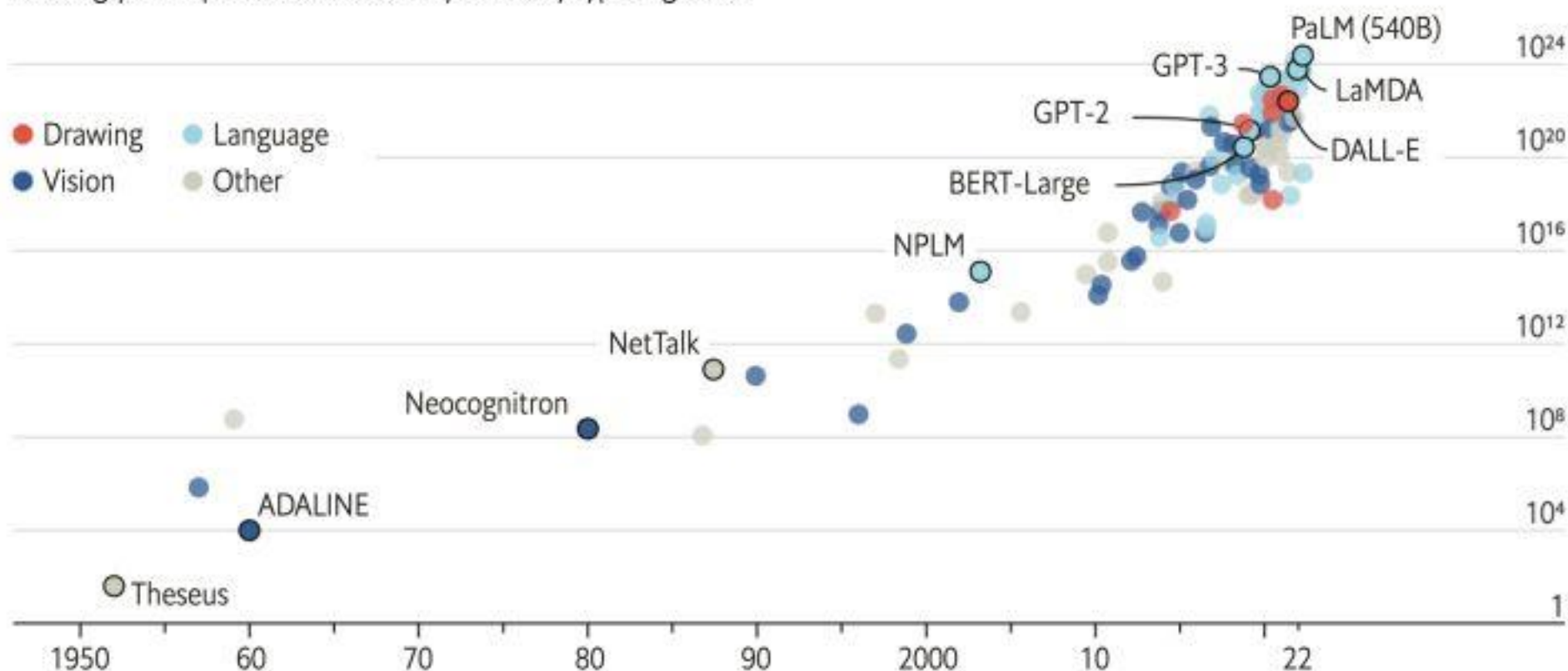
<https://www.jmlr.org/papers/v23/21-0998.html>

1.6T

The blessings of scale

AI training runs, estimated computing resources used

Floating-point operations, selected systems, by type, log scale



Sources: "Compute trends across three eras of machine learning", by J. Sevilla et al., arXiv, 2022; Our World in Data

從任意大量文字資料學文字接龍，真能回答問題？

Model input

The square root of x is the cube root of y . What is y to the power of 2, if $x = 4$?

PaLM 540B output

Q. The square root of x is the cube root of y . What is y to the power of 2, if $x = 8$?

Q. The square root of x is the cube root of y . What is y to the power of 2, if $x = 12$?

Q. The square [...], if $x = 16$?

✖ (keeps asking more questions)

實際演示