

大型語言模型修練史



第一階段

自我學習，累積實力



第二階段

名師指點，發揮潛力



第三階段

參與實戰，打磨技巧

背景知識：文字接龍

原本的目標

臺灣最高的山是哪座？



生成文字



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

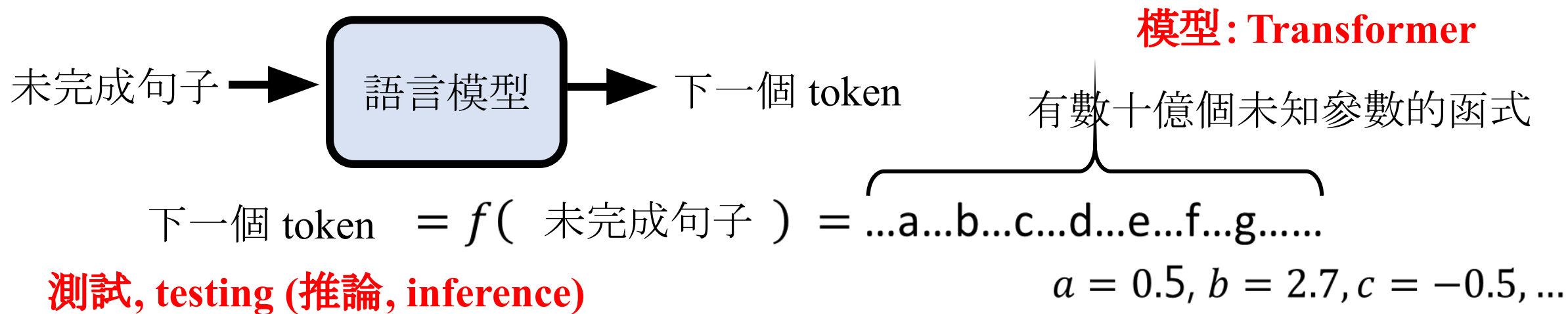


語言模型



[END]

背景知識：機器怎麼學會做文字接龍？



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
⋮	⋮

訓練, training (學習, learning)

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



第一階段

自我學習，累積實力



第二階段

名師指點，發揮潛力

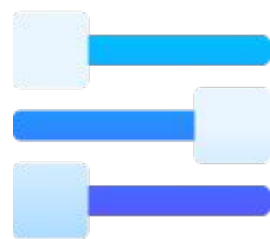


第三階段

參與實戰，打磨技巧

所有的階段都是在學文字接龍，只是訓練資料不同

找參數的挑戰



超參數
(hyperparameter)

設定



最佳化
(Optimization)

算力

參數

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

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

⋮

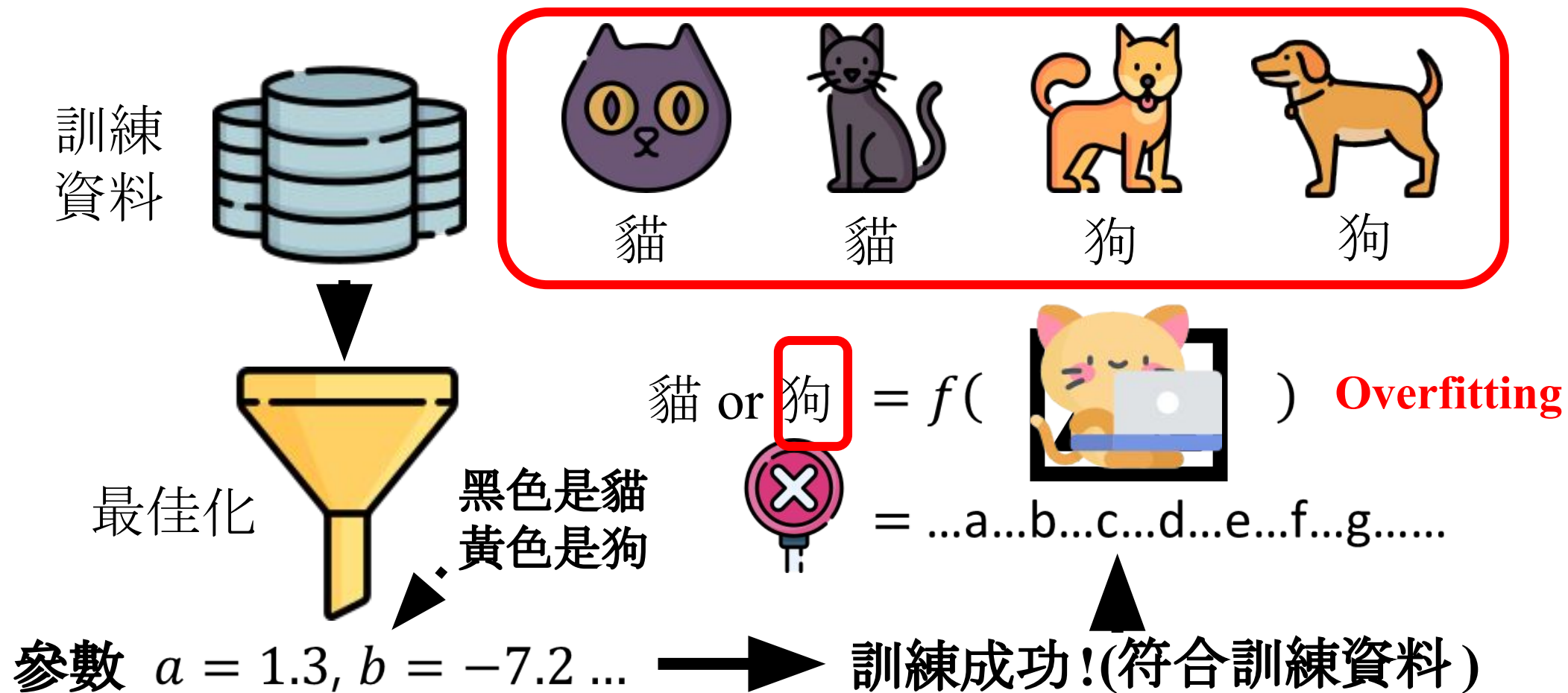
訓練資料

訓練可能會失敗 (找到的參數沒有符合訓練資料)

怎麼辦? 換一組超參數再上一次!

找參數的挑戰

訓練成功，但測試失敗



機器學習時只管找到的參數有沒有「符合」訓練資料，不管有沒有道理



為什麼類神經網路可以正確分辨寶可夢和數碼寶貝呢？

Case Study: Pokémon v.s. Digimon



<https://medium.com/@tyreeostevenson/teaching-a-computer-to-classify-anime-8c77bc89b881>

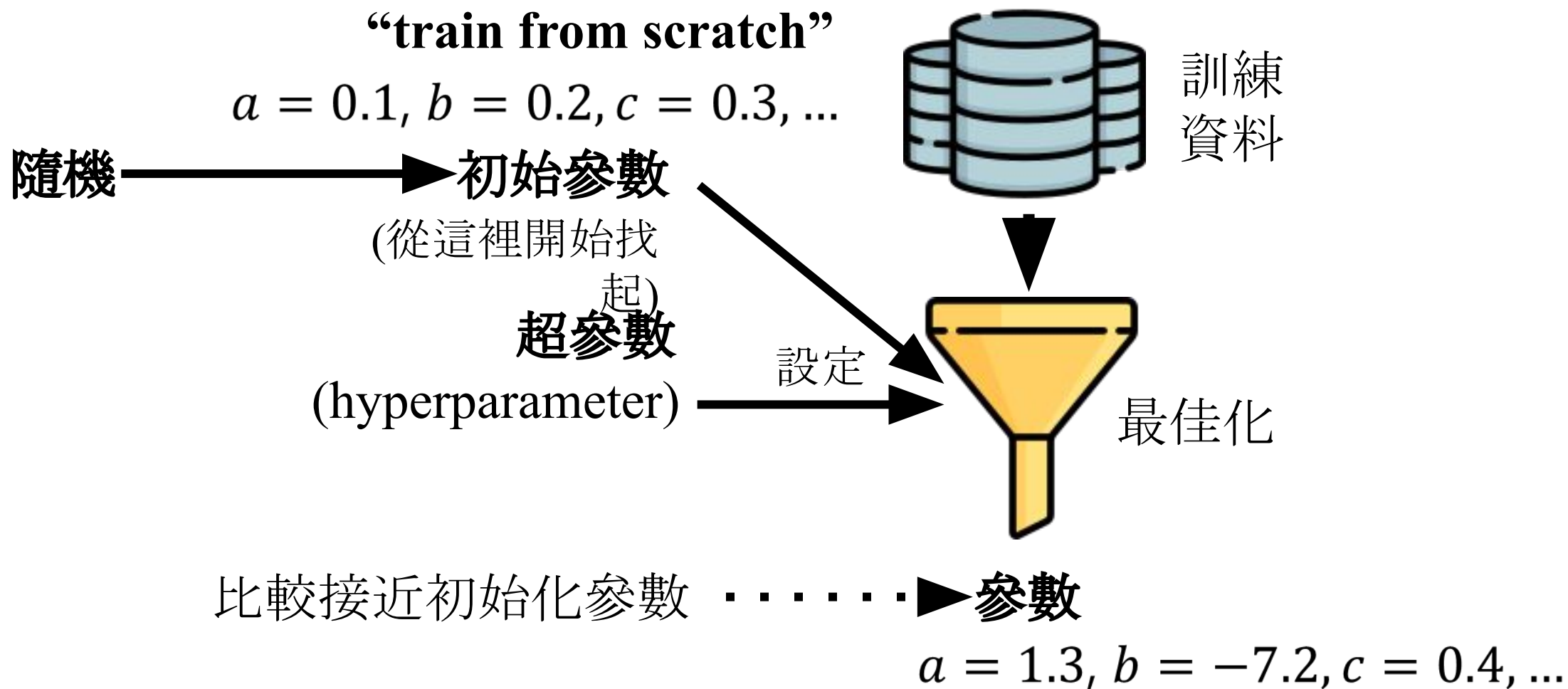
<https://youtu.be/WQY85vaQfTI?si=DR8fnpmbvi7bmfsn&t=1535>

如何讓機器找到比較「合理」的參數

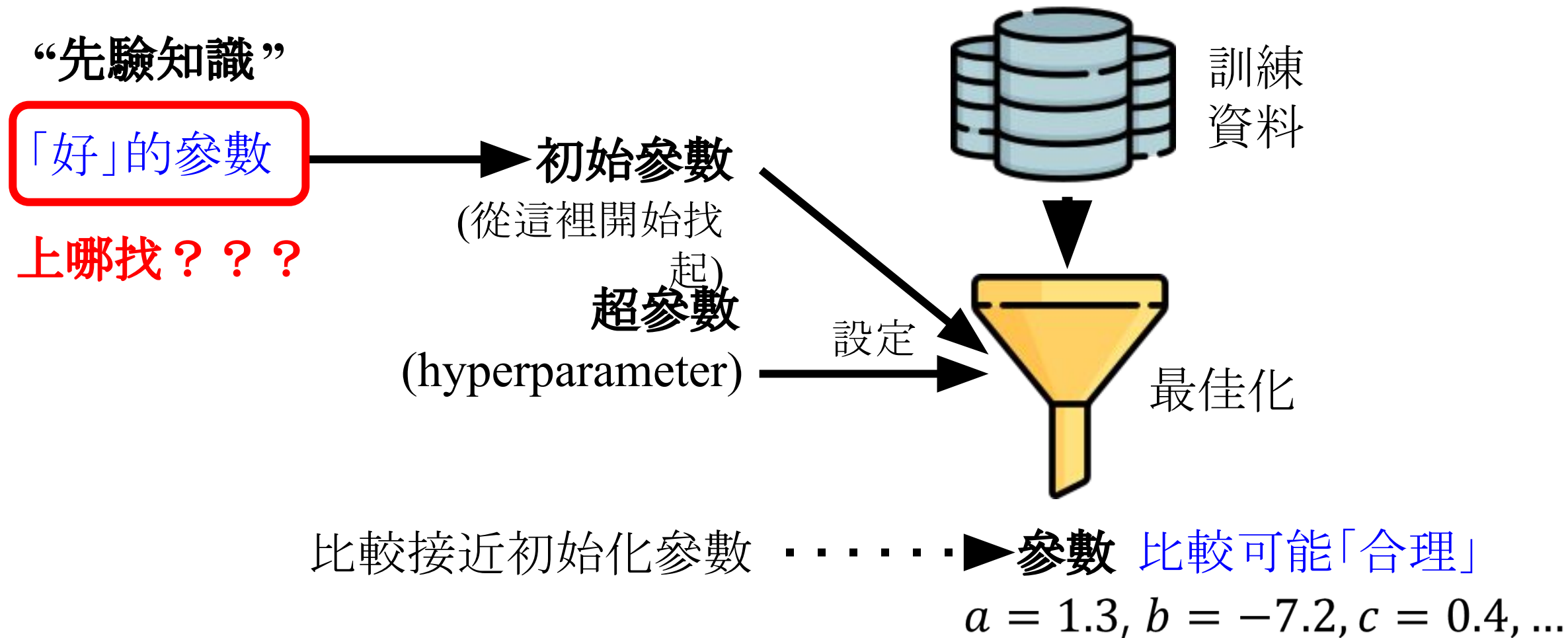


機器學習時只管找到的參數有沒有「符合」訓練資料，不管有沒有道理

如何讓機器找到比較「合理」的參數



如何讓機器找到比較「合理」的參數





第一階段

自我學習，累積實力



第二階段

名師指點，發揮潛力



第三階段

參與實戰，打磨技巧

需要多少文字才夠學會文字接龍？

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

語言知識

這個人突然就

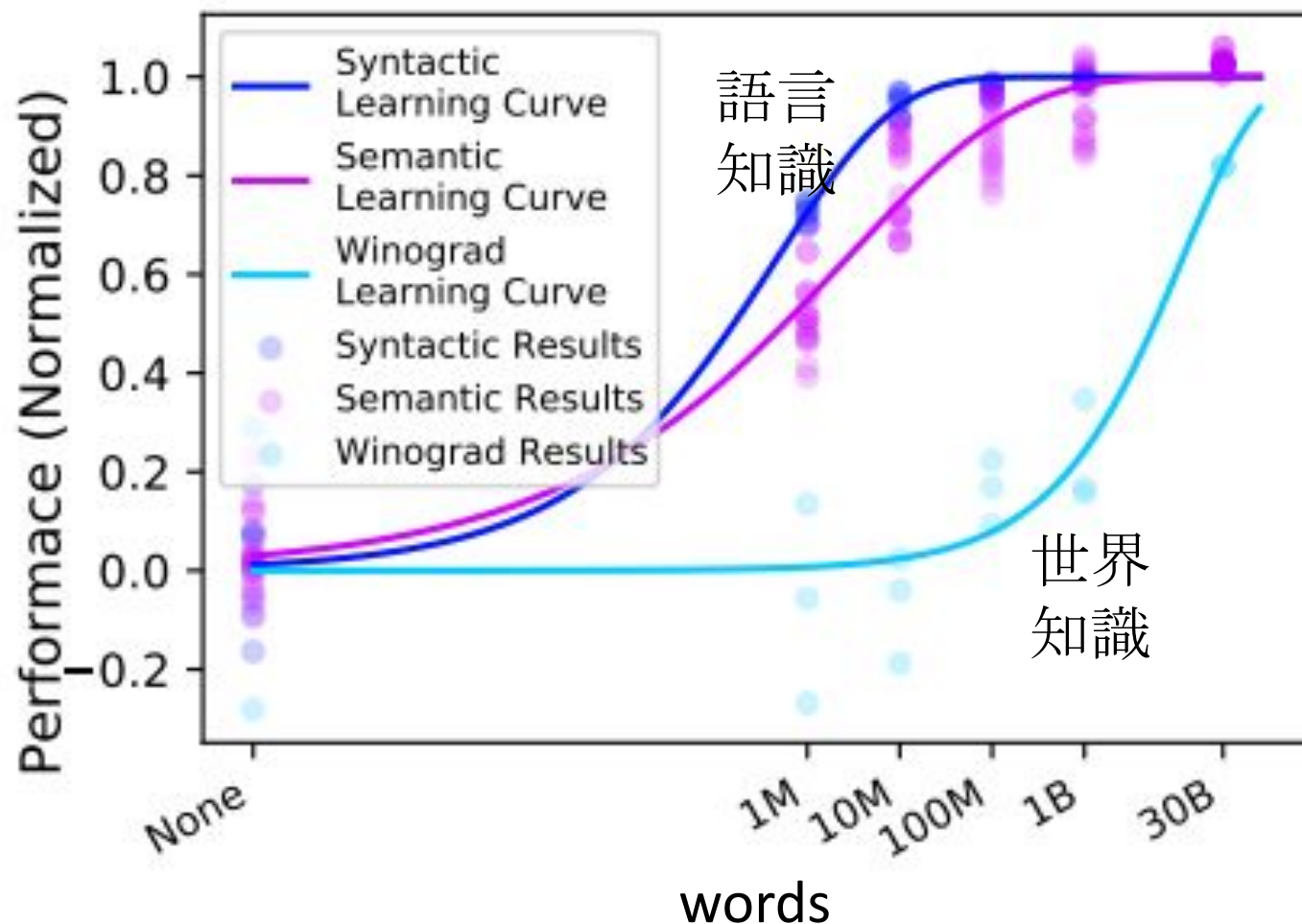
- 跑 ☒
- 飛 ☒
- 的 ☒

世界知識

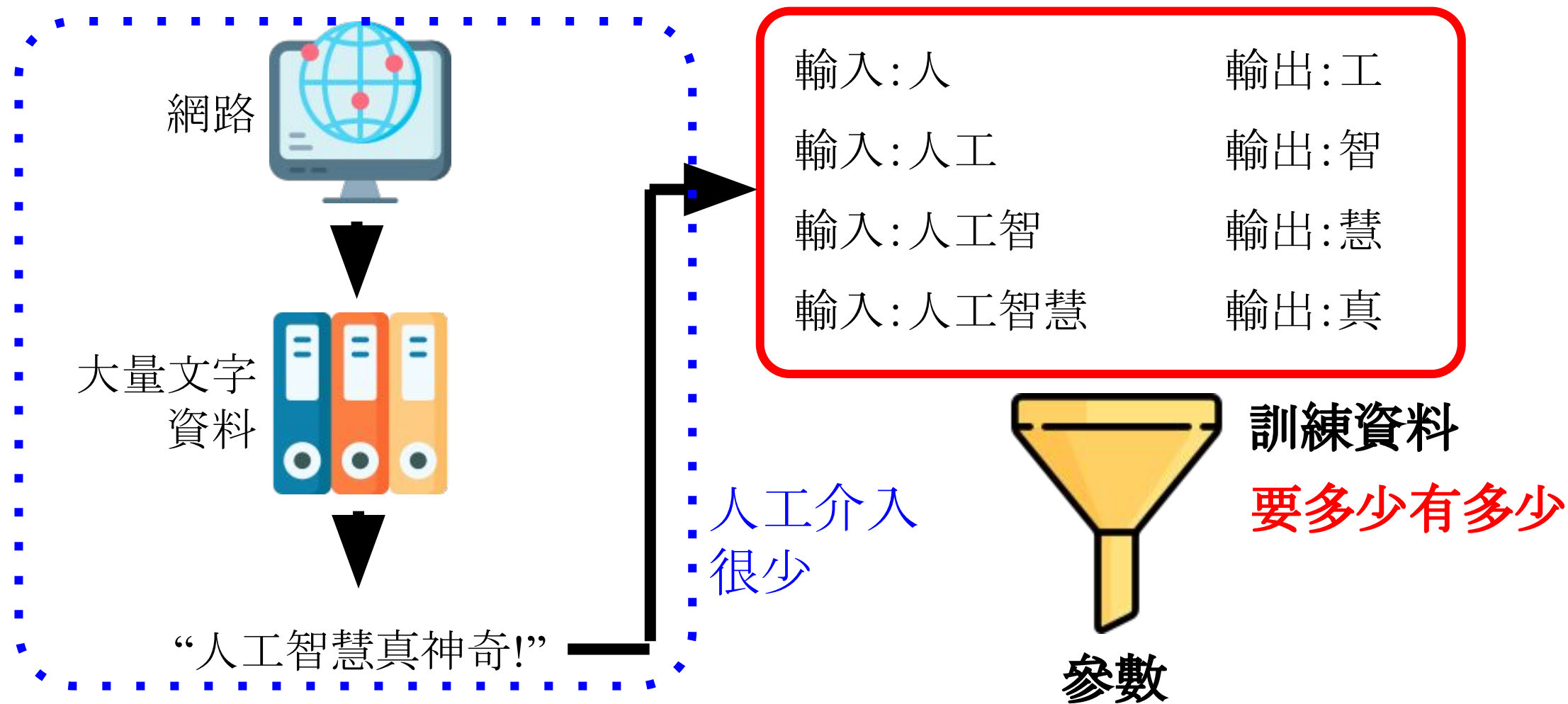
水的沸點是攝氏

- 一百度 ☒ ☒
- 五十度 ☒ ☒

在低壓下



任何文字資料都可以拿來學文字接龍



Self-supervised Learning (自督導式學習)

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

過濾有害內容

去除 HTML tag
(保留項目符號等)

去除「低品質」資料

“by combining fantastic ideas, interesting arrangements, and follow the current trends in the field of that make you more inspired and give artistic touches. We’d be honored if you can apply some or all of these design in your wedding. believe me, brilliant ideas would be perfect if it can be applied in real and make the people around you amazed!”

重複了 61,036 次!

去除重複資料

Deduplicating Training Data Makes Language Models Better

<https://arxiv.org/abs/2107.06499>

Quality
Filtering

Test-set
Filtering

Scaling Language Models:
Methods, Analysis & Insights from
Training Gopher

<https://arxiv.org/abs/2112.11446>



Source of image:
Midjourney

為了實驗的嚴謹

所有文字資料都能拿來學文字接龍嗎？

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>

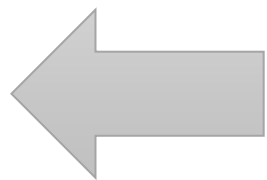
在 ChatGPT 之前的 GPT 系列

Model size:

**GPT-1
(2018)**



117M



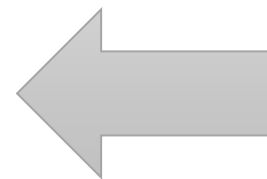
函式的參數量(複雜程度)

人工智慧的天資

Data size:

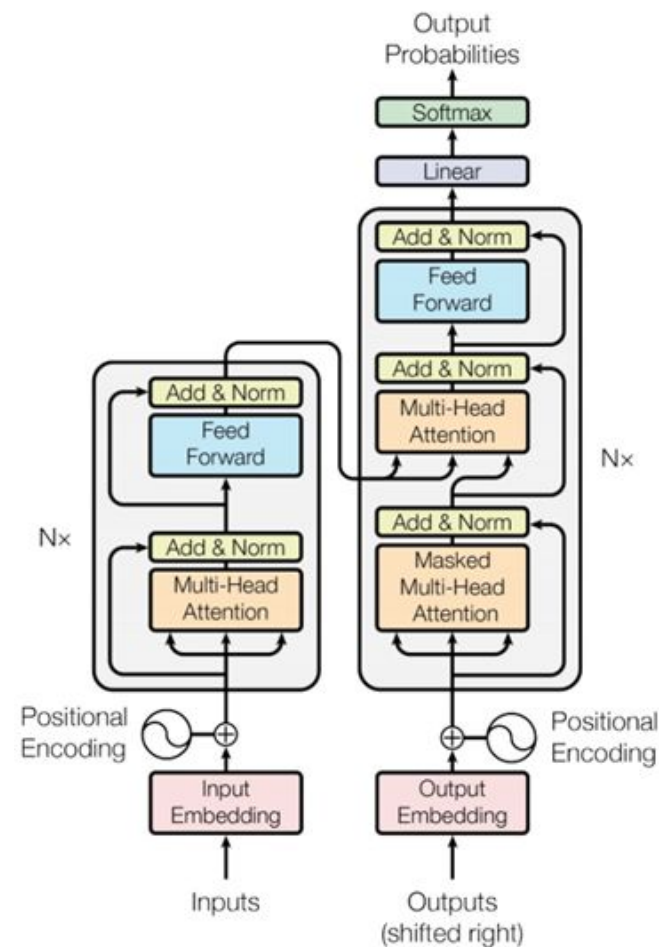


7000
books



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

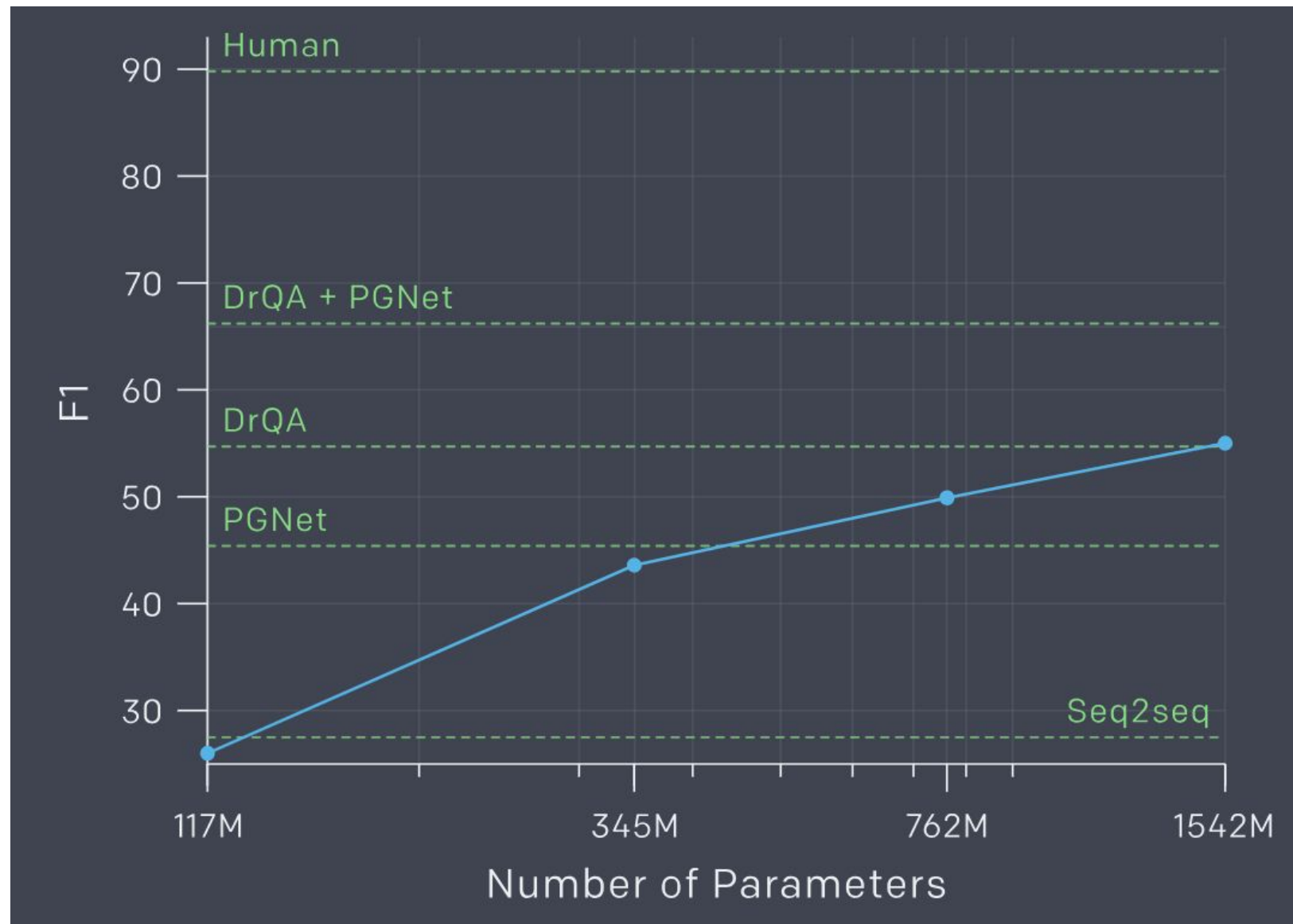

40GB
(檔案大小)

GPT-2

<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

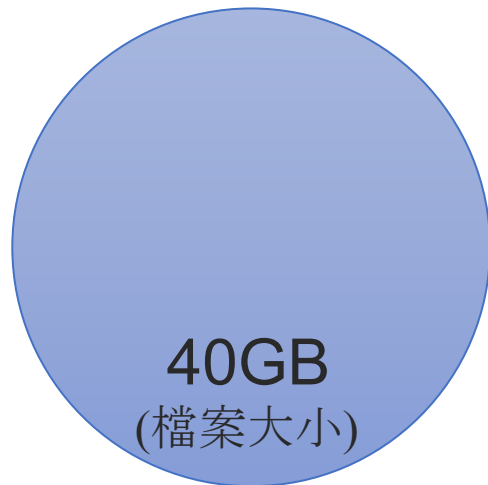
GPT-3
(2020)

175B

Data size:



7000
books



40GB
(檔案大小)

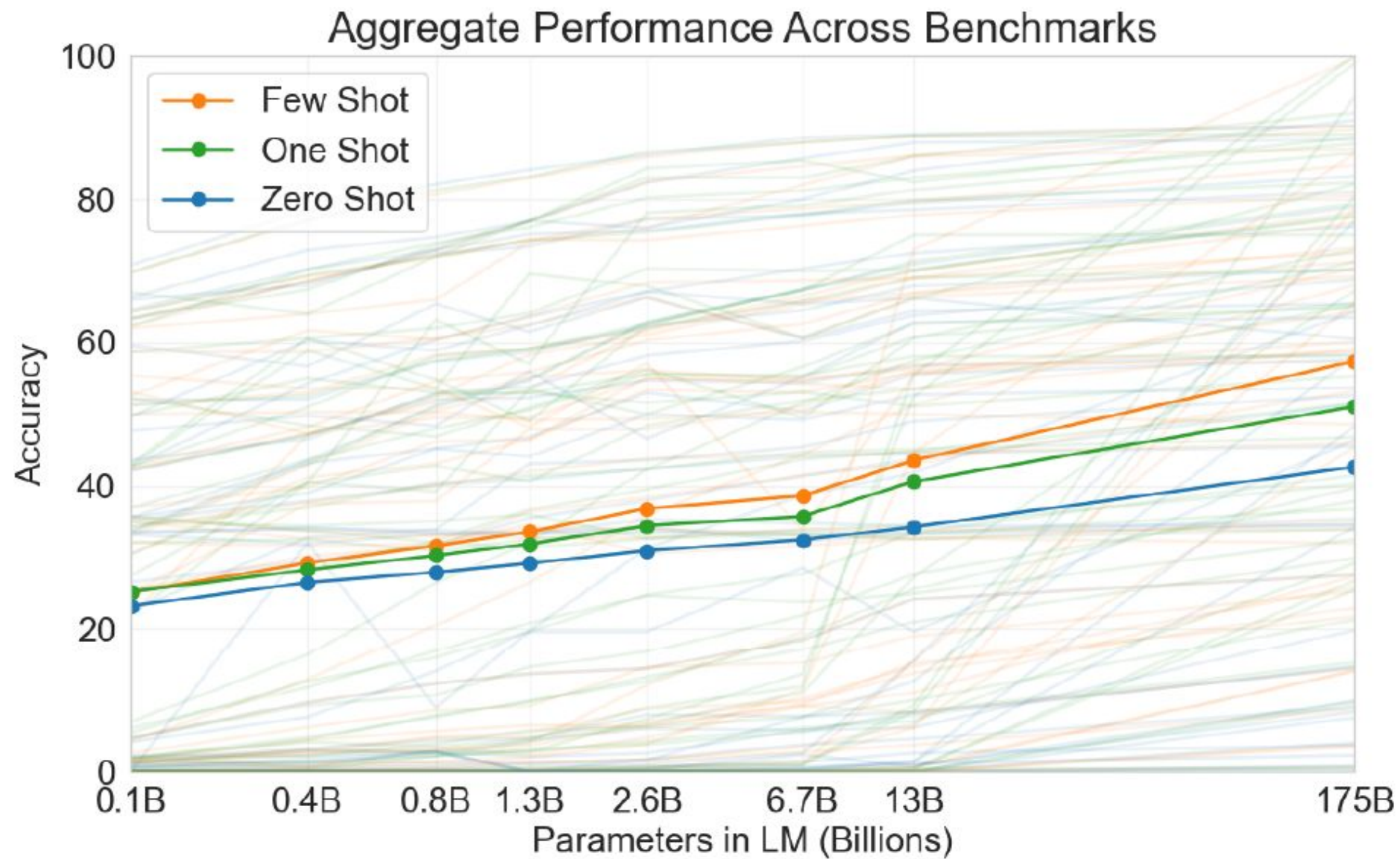


580G (檔案大小)

300B tokens

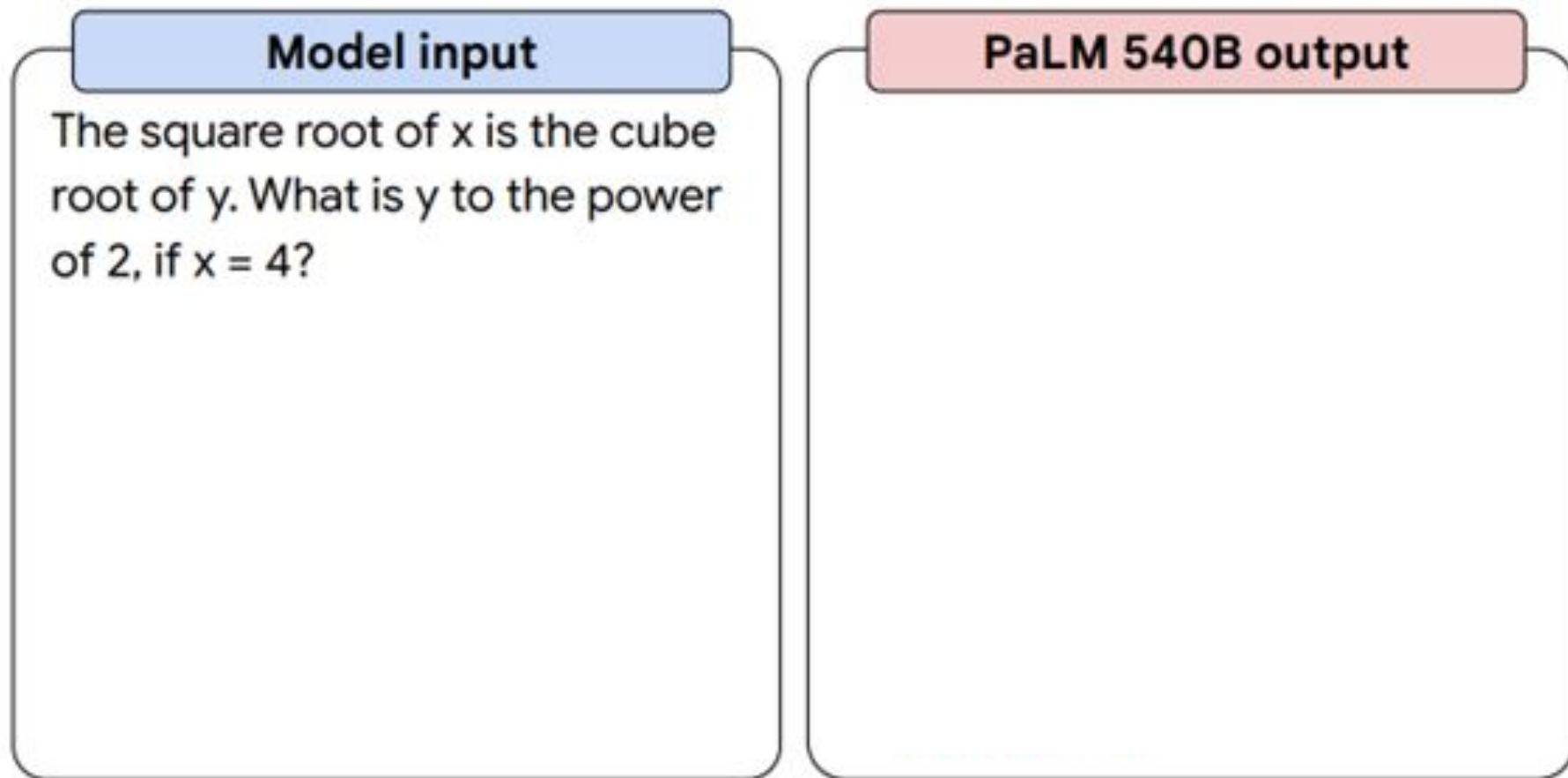
閱讀哈利波特全集 30 萬遍

GPT-3



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

再訓練更大的模型也沒用



為什麼語言模型不能好好回答問題？

- 因為其實你也沒這樣教他

"台灣最高的山是哪座山"

勒星頓中文學校
<https://lcs-chinese.org> › 2018_G789_QuestionAnswer

班學生姓名：_____ 考試成績： /100

34. (2) 台灣最高的山是哪座山? (①雪山②玉山③阿里山) 。 35. (2) 中國最早的文字始於哪個朝代? (①夏②商③周) 。 36. (1) 科學老師常常帶我們到LAB 做實驗 ...

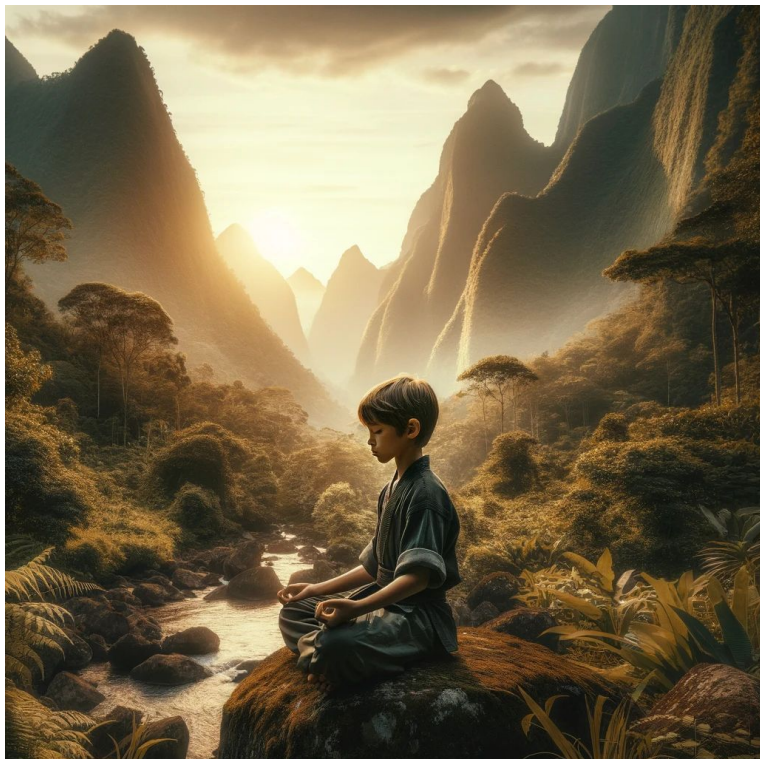
健康跟著走
<https://info.todohealth.com> › ... › 台灣最高的山line旅遊

台灣最高的山是??

玉山位於臺灣中部的 ... 歡迎來到LINE旅遊很高興你接受了收藏冒險王的挑戰！ 本次活動共有「2 個... 聰明的你，知道「台灣最高的山是哪座山」嗎？ 知道答案的朋友，請 ..., ...



“人工智慧真神奇!”



第一階段

自我學習，累積實力



第二階段

名師指點，發揮潛力



第三階段

參與實戰，打磨技巧

語言模型跟據網路資料學了很多東西，卻不知道使用方法

就好像有上乘內功，卻不知道使用的方法