





# 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

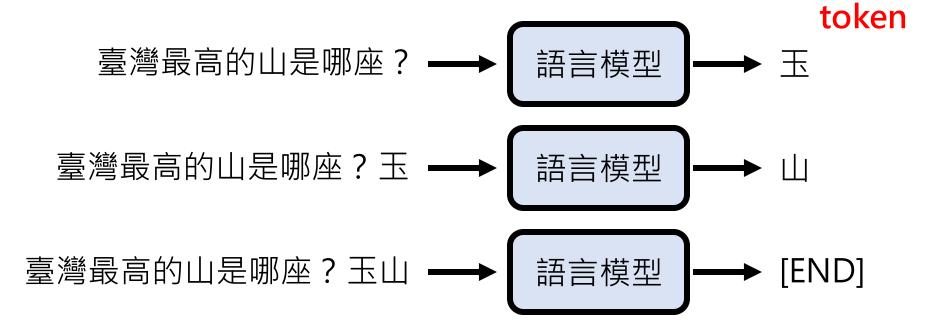


第三階段 參與實戰,打磨技巧

### 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



#### 背景知識:機器怎麼學會做文字接龍?

未完成句子 —— 語言模型 —— 下一個 token

模型:Transformer

有數十億個未知參數的函式

下一個 token = f( 未完成句子 ) = ...a...b...c...d...e...f...g.....

測試, testing (推論, inference)

a = 0.5, b = 2.7, c = -0.5, ...

訓練資料

輸入:人工智

is L

輸入:不要忘了今天來開

輸出:會

輸出:慧

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, ...



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

輸入:人工智

輸入:不要忘了今天來開

輸入:床前明月

設定

•

輸出:慧

輸出:會

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)

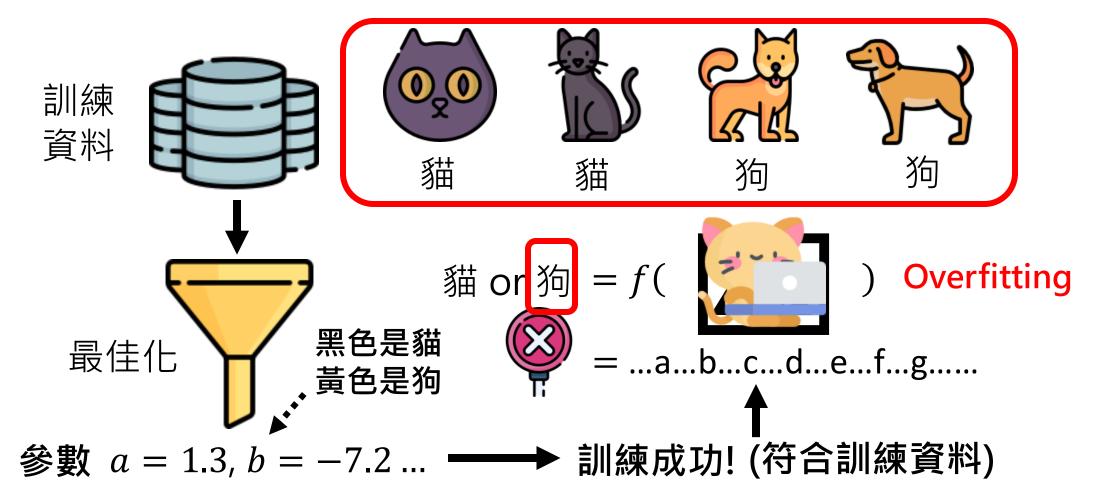
怎麼辦?換一組超參數再上一次!



參數

$$a = 1.3, b = -7.2, c = 0.4, ...$$

## 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



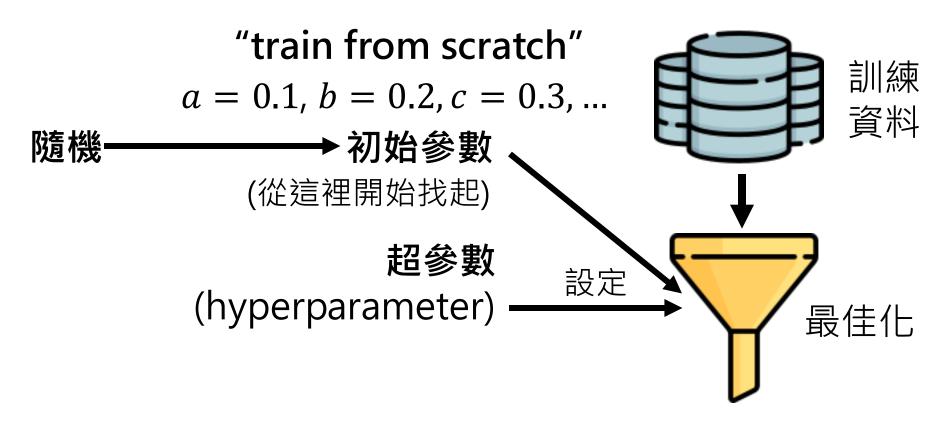
https://youtu.be/WQY85vaQfTI?si=DR8fnpmbvi7bmfsn&t=1535

## 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

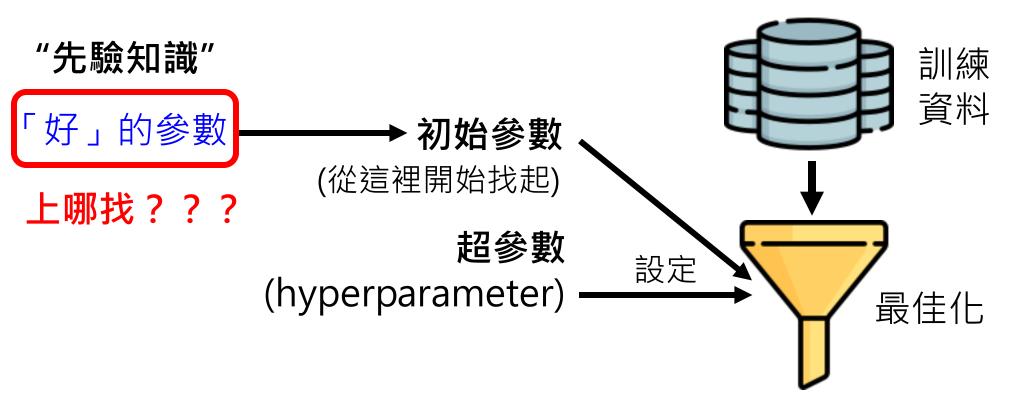
## 如何讓機器找到比較「合理」的參數



比較接近初始化參數 \*\*\*\*\*\* 参數

a = 1.3, b = -7.2, c = 0.4, ...

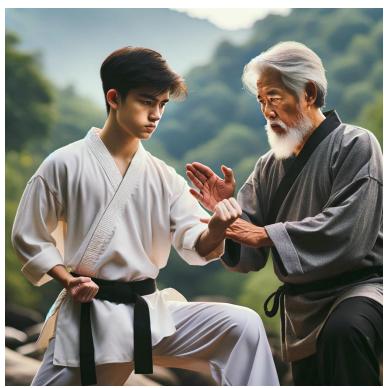
### 如何讓機器找到比較「合理」的參數



比較接近初始化參數 •••••• ◆數 比較可能「合理」 a = 1.3, b = -7.2, c = 0.4, ...



第一階段 自我學習,累積實力



第二階段 名師指點,發揮潛力

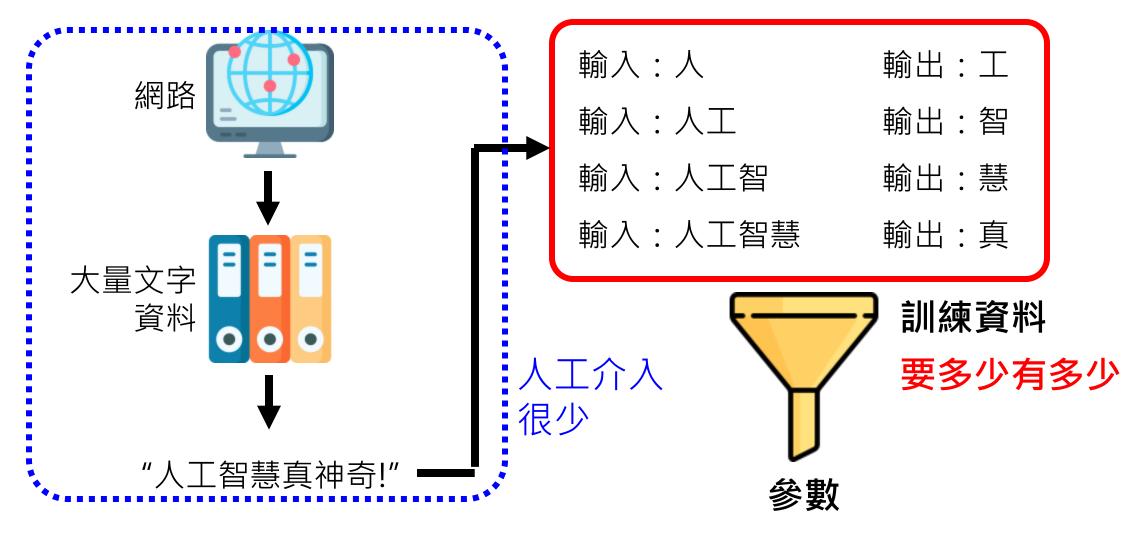


第三階段 參與實戰,打磨技巧

### 需要多少文字才夠學會文字接龍?

https://arxiv.org/abs/2011.04946 語言知識 Syntactic 1.0 語言 跑 Learning Curve (Normalized) Semantic 知識 0.8 這個人突然就 Learning Curve Winograd 0.6 Learning Curve Syntactic Results 0.4Semantic Results Winograd Results 世界知識 百度 世界 知識 水的沸點是摄氏 在低壓下 五十度 words

## 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

過濾有害內容

去除 HTML tag (保留項目符號等)

"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!

• 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

ality

ering

t-set

ering

去除「低品質」資料

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

https://arxiv.org/abs/2112.11446



Source of image: Midjourney

去除重複資料

Deduplicating Training Data Makes Language Models Better https://arxiv.org/abs/2107.06499 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

# OpenAl and journalism

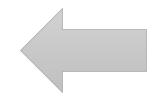
We support journalism, partner with news organizations, and believe The New York Times lawsuit is without merit.

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

#### 在 ChatGPT 之前的 GPT 系列 ......

Model size:





函式的參數量(複雜程度)

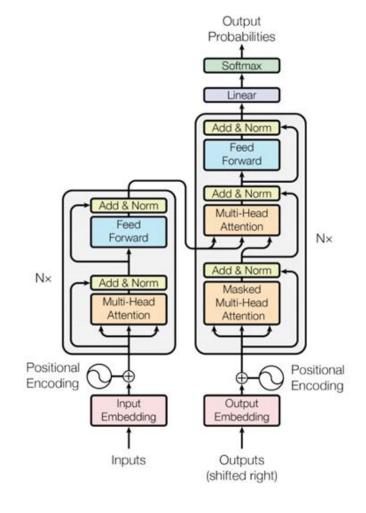
人工智慧的天資

Data size:

books

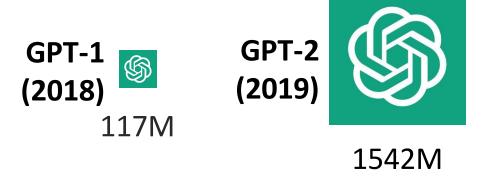


拿來學文字接龍的資料量 後天的努力



#### 在 ChatGPT 之前的 GPT 系列 ......

#### Model size:



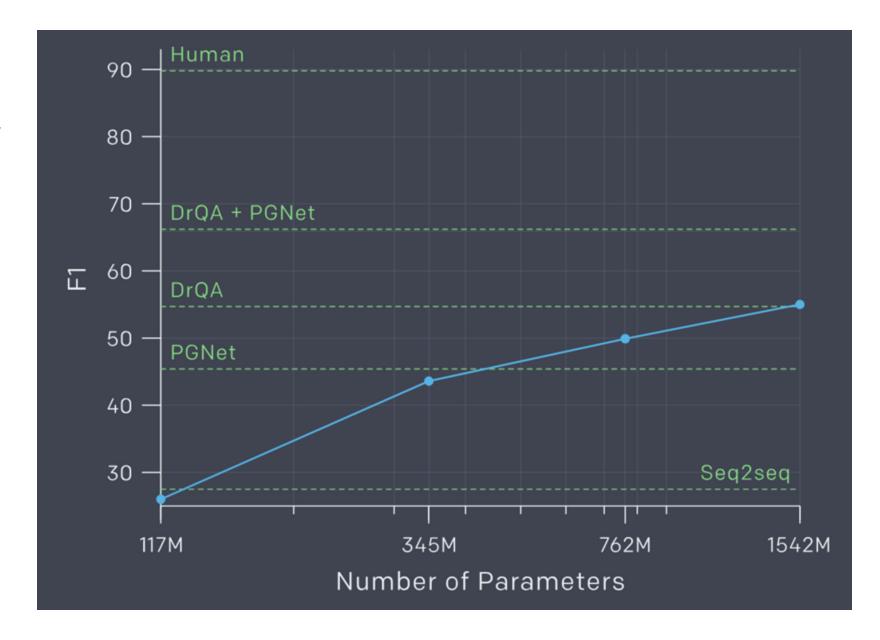
Data size:



#### 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-3 (2020)

175B

Data size:

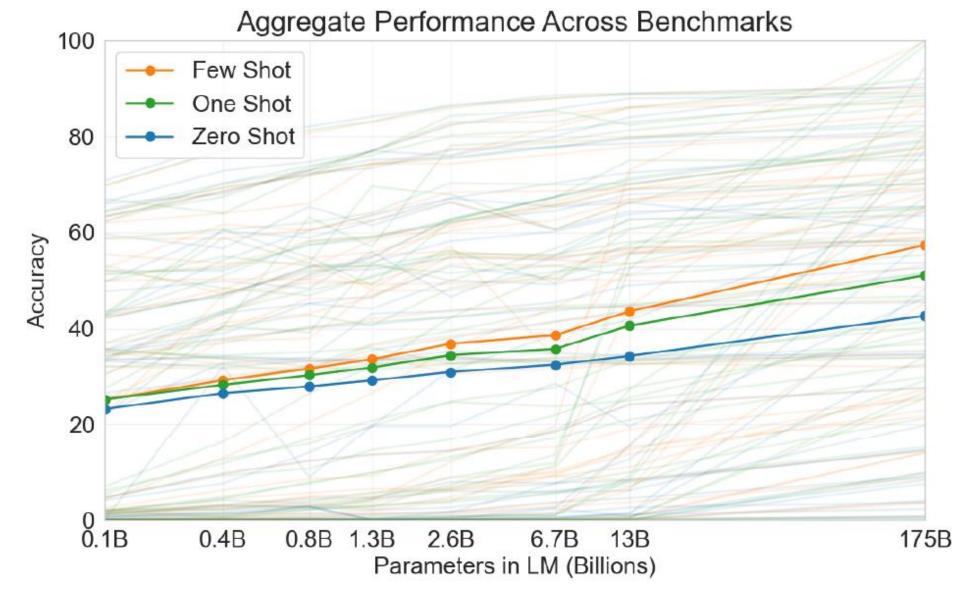


**40GB** (檔案大小) 580GB (檔案大小)

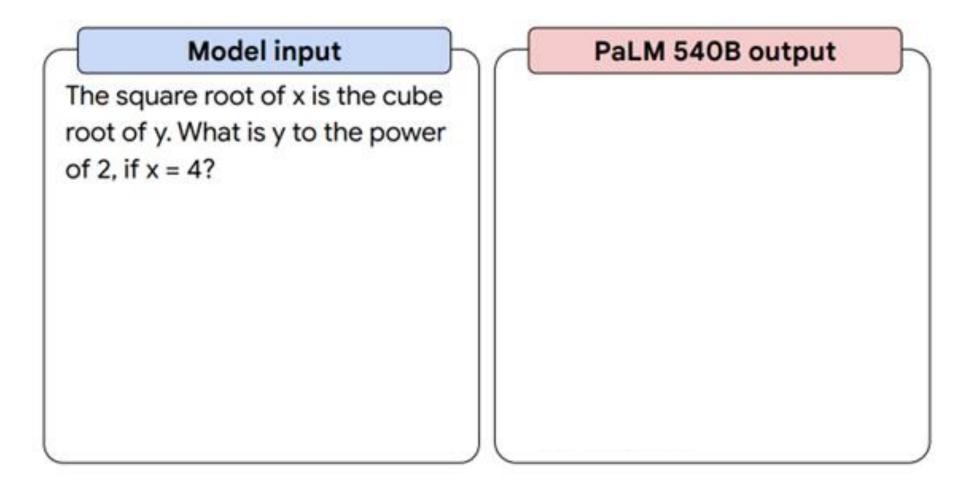
300B tokens

閱讀哈利波特全集 30 萬遍





#### 再訓練更大的模型也沒用.....



https://ai.googleblog.com/2022/11/better-language-models-without-massive.html

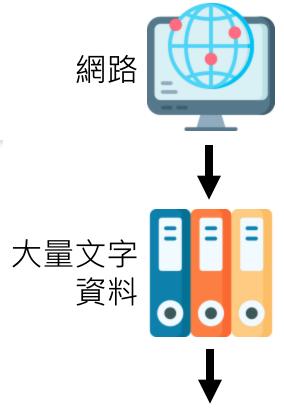
## 為什麼語言模型不能好好回答問題?

• 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??

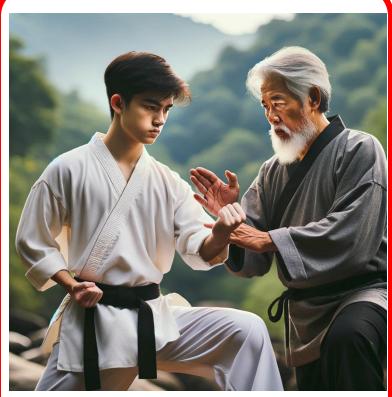
玉山位於臺灣中部的 ... 歡迎來到LINE旅遊很高興你接受了收藏冒險王的挑戰! 本次活動共有「2 個... 聰明的你,知道「台灣最高的山是哪座山」嗎? 知道答案的朋友,請 ........



"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