語言模型運作

黃志勝（Tommy Huang）
義隆電子 人工智慧研發部
國立陽明交通大學 AI學院 合聘助理教授
國立台北科技大學 電資學院合聘助理教授

大家有沒有想過知識學習的程序是什麼

大家有沒有想過知識學習的程序是什麼

人之初
性本善
性相近
習相遠

人之初
性本善
性相近
習相遠

小朋友多唸幾次

大家有沒有想過知識學習的程序是什麼

Language Modeling

基於token序列的機率分布 $p(x_1, x_2, \ldots, x_L)$

$$p(人, 之, 初, 性, 本, 善) = 0.8$$
$$p(人, 性, 本, 善, 初, 之) = 0.1$$
$$p(性, 本, 善, 人, 之, 初) = 0.01$$

LMs are generative models

$$x_i \sim p(x_1, x_2, \ldots, x_L), \forall i = 1,2, \ldots, L$$

Autoregressive (AR) language

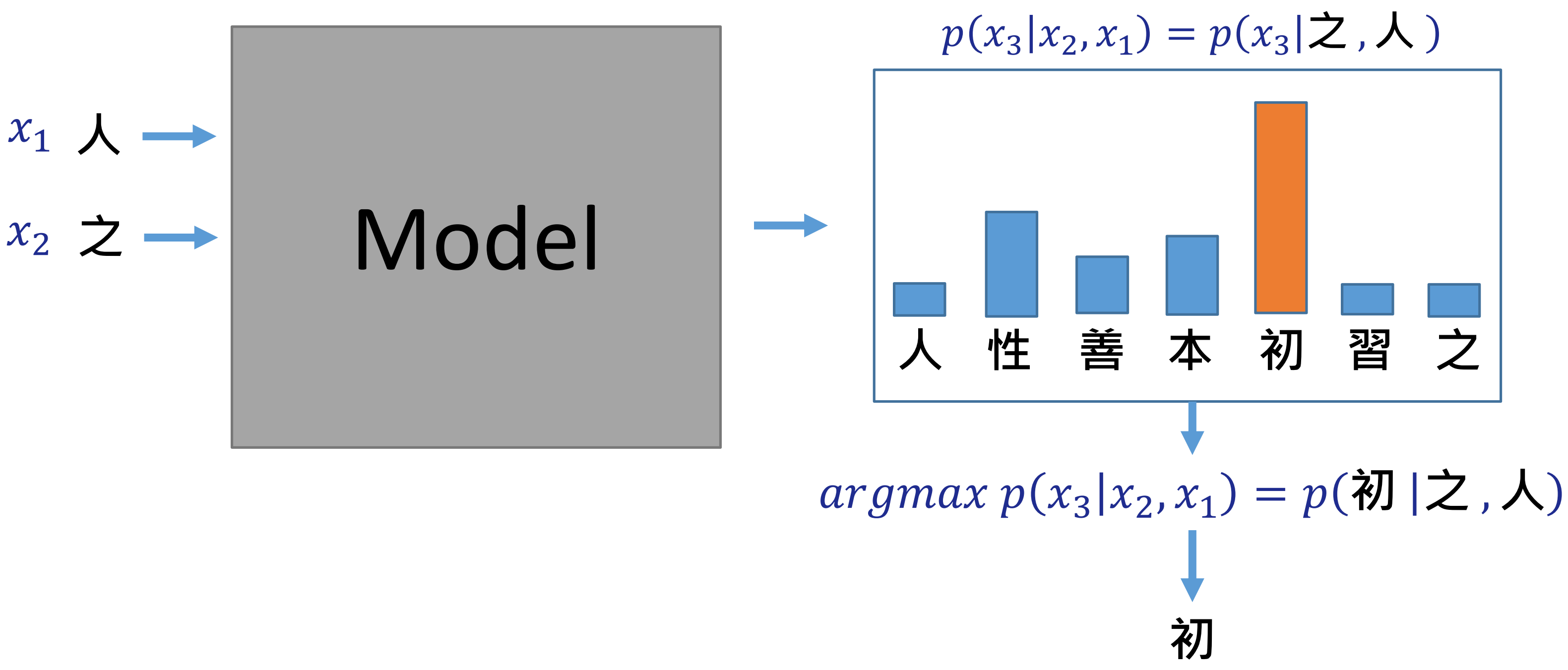
$$p(x_1, x_2, \dots, x_L) = p(x_1)p(x_2|x_1)p(x_3|x_2, x_1) \dots = \prod_{i=1}^{L} p(x_i|x_1, \dots, x_{i-1})$$

看不懂沒關係

簡單說你需要一個模型基於**過去的內容**預測**下一次的Token**。

Autoregressive (AR) model

任務: 預測下一個字



x_1 人

x_2 之

Model

$p(x_3|x_2,x_1) = p(x_3|$之,人$)$

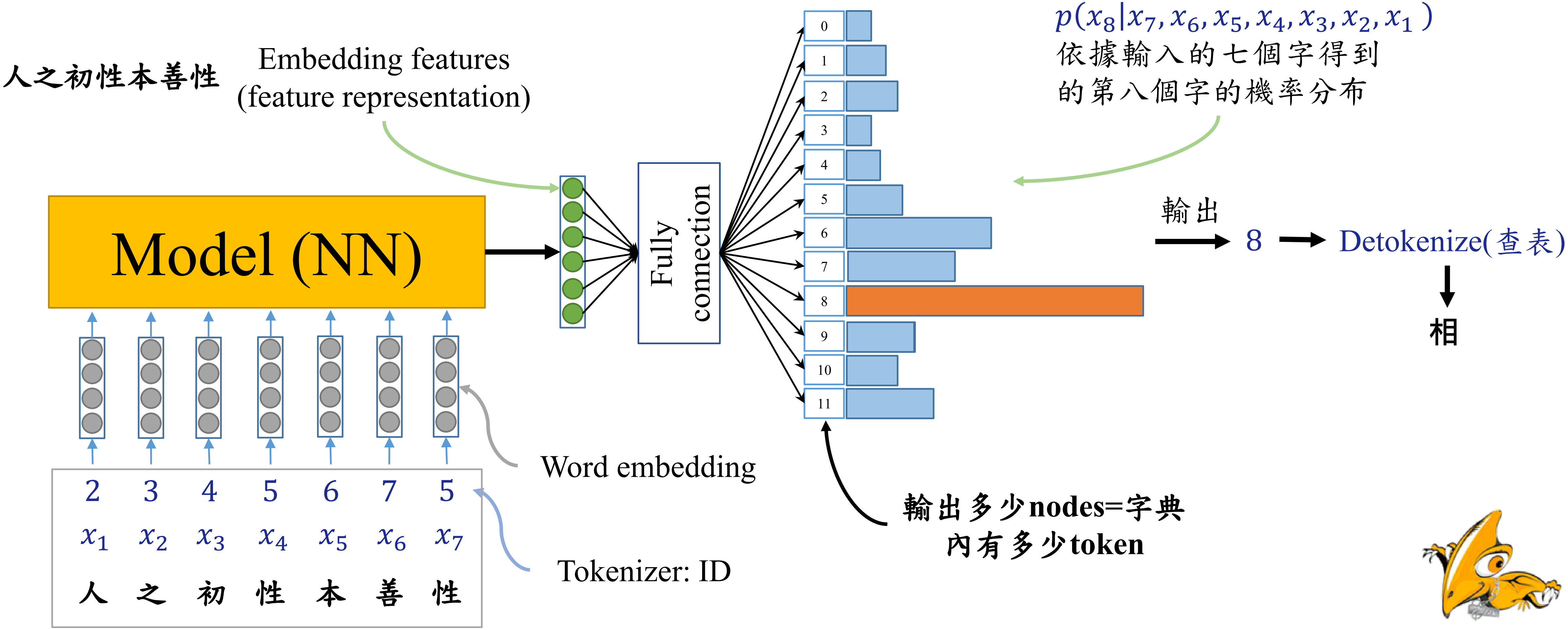
$argmax\ p(x_3|x_2,x_1) = p($初$|$之,人$)$

初

https://lena-voita.github.io/nlp_course/language_modeling.html#intro

AR Neural Language Models

	人	之	初	性	本	善	相	近	習	遠	EOS	Pad	遠
ID	2	3	4	5	6	7	8	9	10	11	1	0	12



人之初性本善性

Embedding features (feature representation)

$p(x_8|x_7, x_6, x_5, x_4, x_3, x_2, x_1)$

依據輸入的七個字得到的第八個字的機率分布

Word embedding

Tokenizer: ID

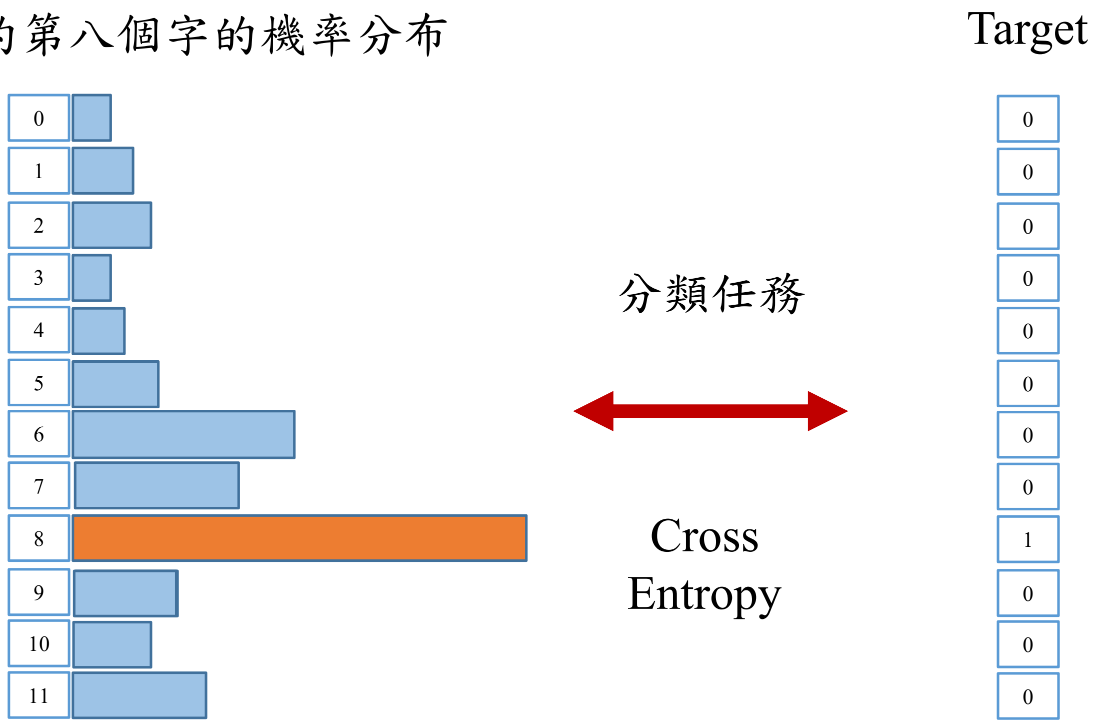
輸出多少nodes=字典內有多少token

怎麼訓練: Loss

	人	之	初	性	本	善	相	近	習	遠	EOS	Pad	遠
ID	2	3	4	5	6	7	8	9	10	11	1	0	12

$p(x_8|x_7,x_6,x_5,x_4,x_3,x_2,x_1\ )$

依據輸入的七個字得到
的第八個字的機率分布



Context:

人之初性本善性相近習相遠<EOS>

↓

輸入7個字

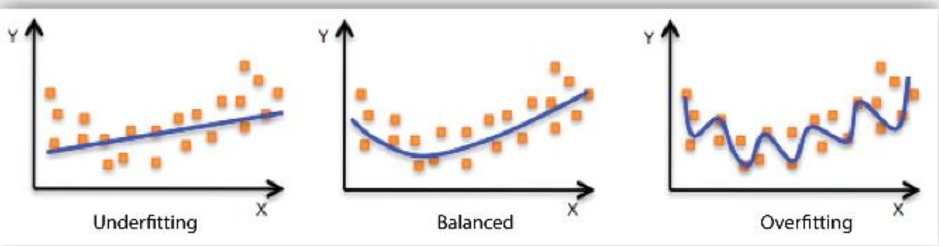
↓

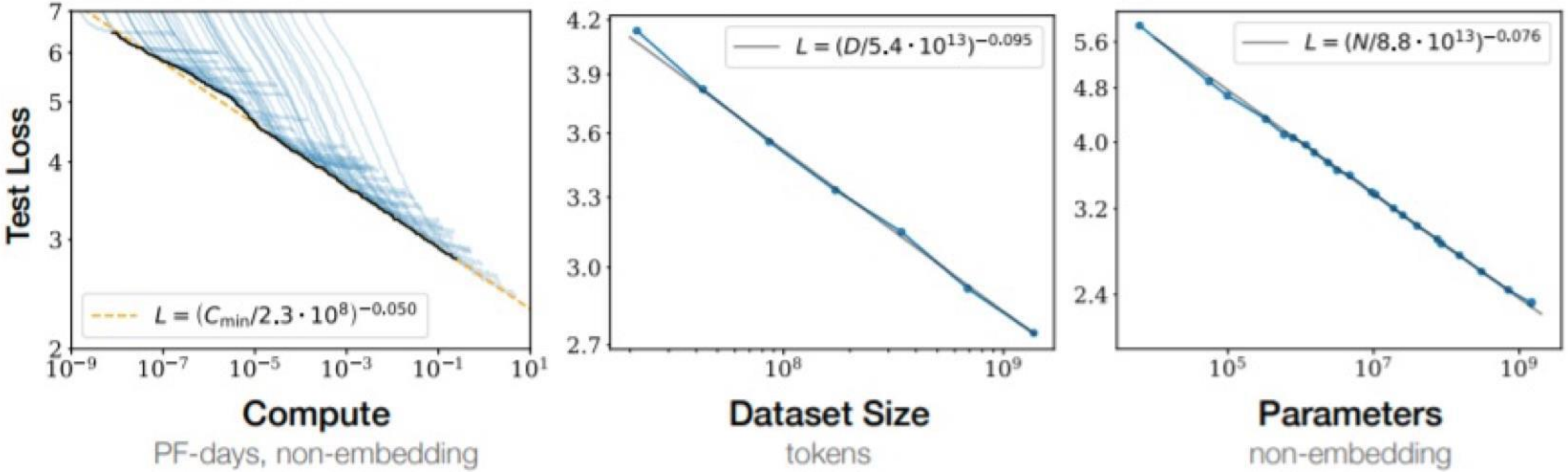
1. 人之初性本善性→相
2. 之初性本善性相→近
3. 初性本善性相近→習
4. 性本善性相近習→相
5. 本善性相近習相→遠
6. 善性相近習相遠→<EOS>

Scaling laws

Scaling laws 是機器學習領域中描述模型性能與數據量、模型大小、計算資源之間關係的一組規則。

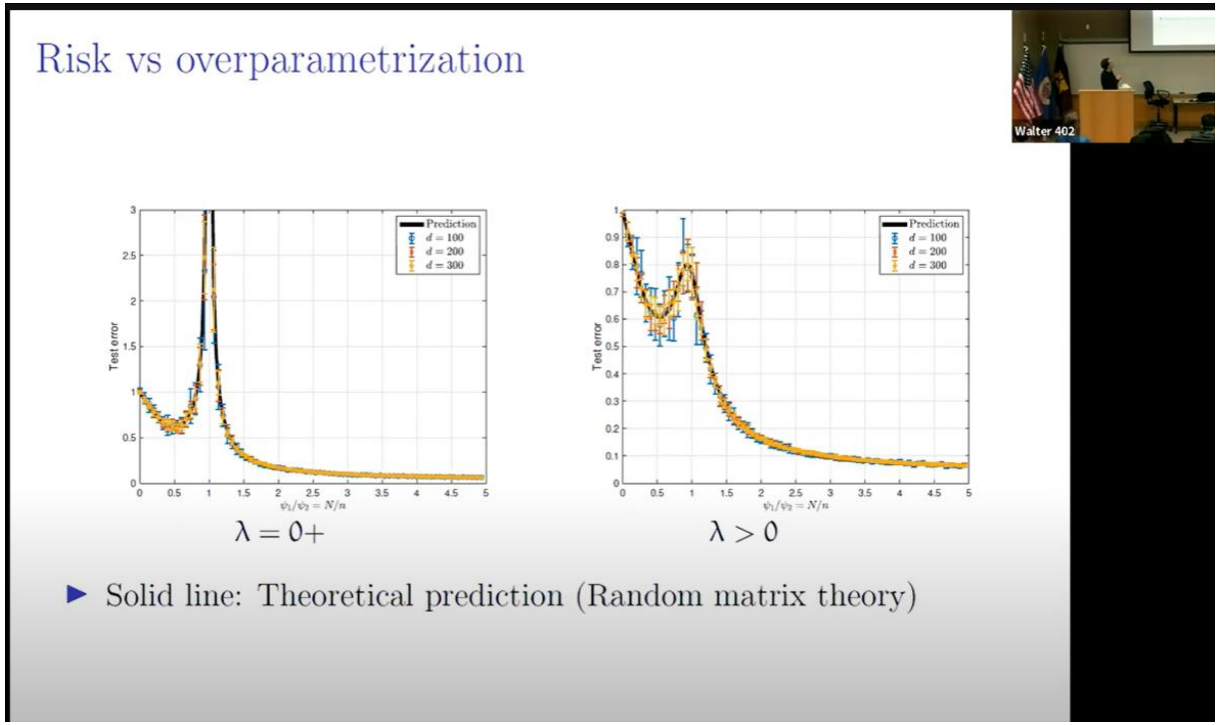
- 通常越多資料越大的模型→表現得越好
- 在深度學習年代: 大模型不代表會overfitting





Scaling laws [Kaplan+ 2020]

Overparameter



Risk vs overparametrization

$\lambda = 0+$ $\lambda > 0$

- Solid line: Theoretical prediction (Random matrix theory)

https://www.youtube.com/watch?v=FiX-u9hkuo0

我不仔細講

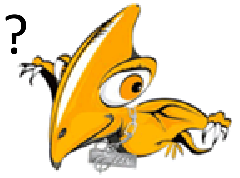
有興趣可以看NTK(neural tangent kernel)/YT連結

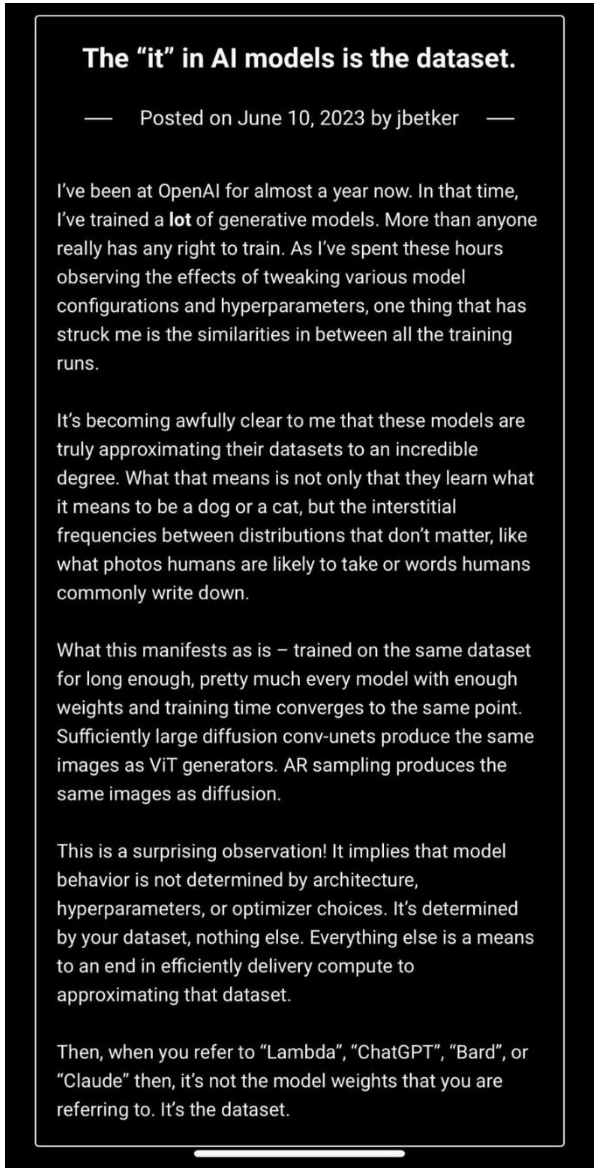
但基本上在現代AI如果發生overfitting的問題，基本上就是data overfitting。

任何超強超大的learning演算法都只是在學習Data的upper bound。

你讓小朋友學習「加法」和「減法」

他有辦法會「乘法」和「除法」嗎?





The “it” in AI models is the dataset.

— Posted on June 10, 2023 by jbetker —

I’ve been at OpenAI for almost a year now. In that time, I’ve trained a **lot** of generative models. More than anyone really has any right to train. As I’ve spent these hours observing the effects of tweaking various model configurations and hyperparameters, one thing that has struck me is the similarities in between all the training runs.

It’s becoming awfully clear to me that these models are truly approximating their datasets to an incredible degree. What that means is not only that they learn what it means to be a dog or a cat, but the interstitial frequencies between distributions that don’t matter, like what photos humans are likely to take or words humans commonly write down.

What this manifests as is – trained on the same dataset for long enough, pretty much every model with enough weights and training time converges to the same point. Sufficiently large diffusion conv-unets produce the same images as ViT generators. AR sampling produces the same images as diffusion.

This is a surprising observation! It implies that model behavior is not determined by architecture, hyperparameters, or optimizer choices. It’s determined by your dataset, nothing else. Everything else is a means to an end in efficiently delivery compute to approximating that dataset.

Then, when you refer to “Lambda”, “ChatGPT”, “Bard”, or “Claude” then, it’s not the model weights that you are referring to. It’s the dataset.

- **"Data is Everything".**

一位OpenAI的員工在2023 年 6 月 10 日 jbetker 發表,大意是:

- **AI 模型的關鍵在於資料集**，而不是其他。
- 模型的行為，並不取決於架構、超參數或優化器的選擇，而是完全由訓練資料決定。其他的一切，都只是為了高效利用算力，去擬合那批資料而已。
- 所以,當我們談論 "Lambda"、"ChatGPT"、"Bard" 或 "Claude" 的時候，指的其實不是模型的權重，而是它們背後的資料集。

怎麼評估(evaluation): Perplexity

Perplexity (PPL) →困惑

什麼情況模型會困惑→不確性高的時候

不確性怎麼評估→Entropy $H(X) = -\sum_i p_i \log(p_i)$

骰子1

出現次數

4							0.333
3							0.250
2	🟩	🟩	🟩	🟩	🟩	🟩	0.167
1	🟩	🟩	🟩	🟩	🟩	🟩	0.083
	1	2	3	4	5	6	

出現機率

出現點數

$$H(X) = -\sum_{i=1}^{6} p_i \log(p_i)$$
$$= -6 * 0.167 * \log(0.167) = 0.779$$

輸出的機率越集中，
模型困惑度越低。

骰子2

出現次數

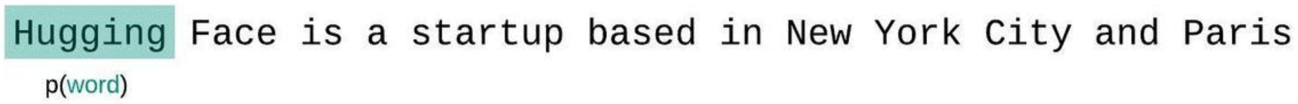
4			🟩	🟩			0.333
3			🟩	🟩			0.250
2			🟩	🟩			0.167
1	🟩	🟩	🟩	🟩	🟩	🟩	0.083
	1	2	3	4	5	6	

出現機率

出現點數

$$H(X) = -\sum_{i=1}^{6} p_i \log(p_i)$$
$$= -4 * 0.083 * \log(0.083) - 2 * 0.333 * \log(0.333)$$
$$= 0.359 + 0.318 = 0.677$$

怎麼評估(evaluation): Perplexity



Hugging Face is a startup based in New York City and Paris

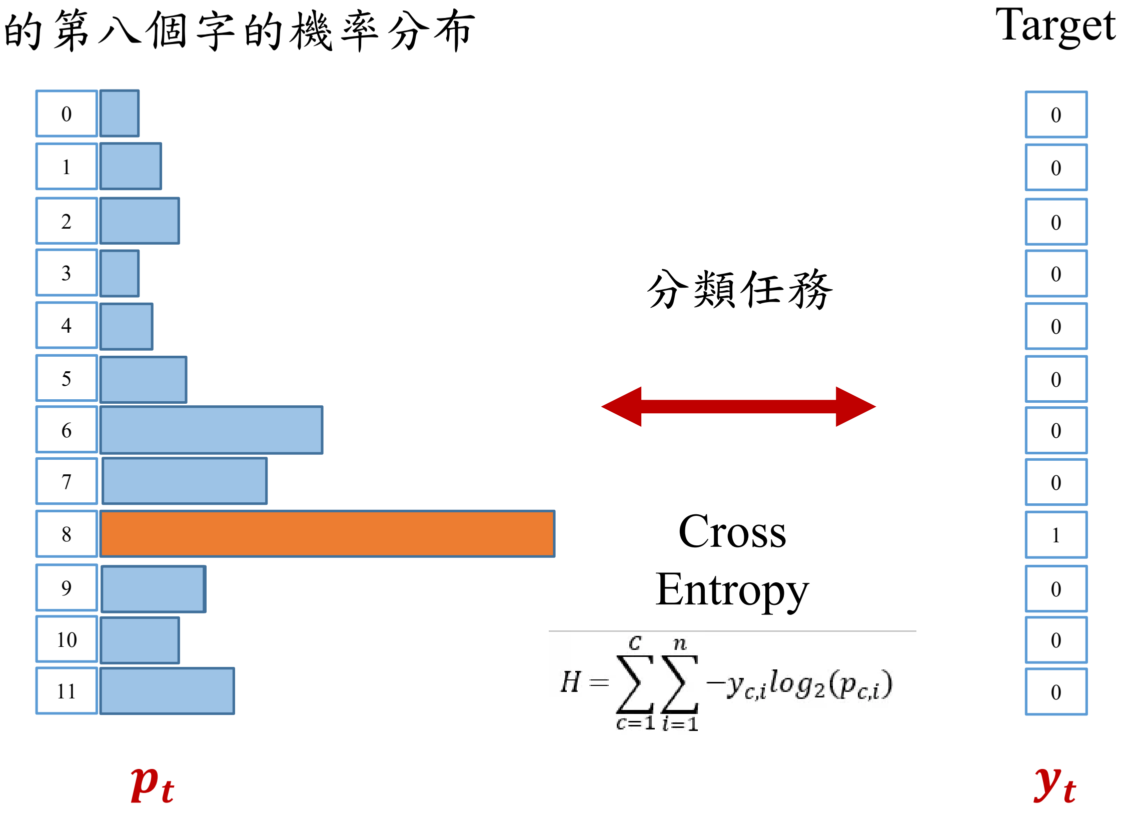
p(word)

https://huggingface.co/docs/transformers/perplexity

$p(x_8|x_7,x_6,x_5,x_4,x_3,x_2,x_1)$

依據輸入的七個字得到

的第八個字的機率分布



0	0
1	0
2	0
3	0
4	0
5	0
6	0
7	0
8	1
9	0
10	0
11	0

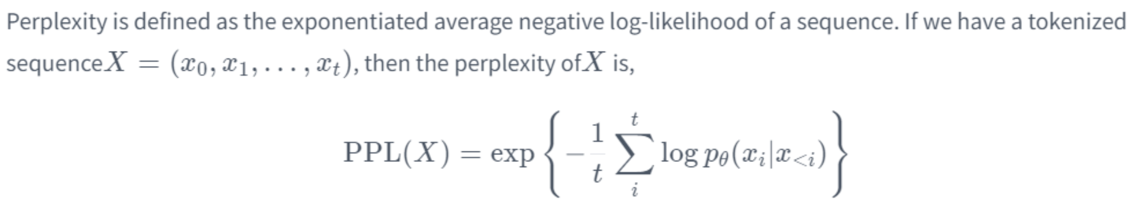
$$H=\sum_{c=1}^{C}\sum_{i=1}^{n}-y_{c,i}log_2(p_{c,i})$$

假設我們只看這筆資料的cross entropy

L: number of vocab

$$H(X)=-\sum_{t=1}^{L}y_t log_2(p_t)$$

https://huggingface.co/docs/transformers/perplexity



Perplexity is defined as the exponentiated average negative log-likelihood of a sequence. If we have a tokenized sequence $X=(x_0,x_1,\ldots,x_t)$, then the perplexity of X is,

$$\text{PPL}(X)=\exp\left\{-\frac{1}{t}\sum_{i}^{t}\log p_\theta(x_i|x_{<i})\right\}$$

$$p(x_1,x_2,\ldots,x_L)=\prod_{i=1}^{L}p(x_i|x_1,\ldots,x_{i-1})$$

$$\Rightarrow PPL(X)=p(x_1,x_2,\ldots,x_L)^{-\frac{1}{L}}=\left(\prod_{i=1}^{L}p(x_i|x_1,\ldots,x_{i-1})\right)^{-\frac{1}{L}}$$

$$\Rightarrow e^{\log\left(p(x_1,x_2,\ldots,x_L)^{-\frac{1}{L}}\right)}=e^{-\frac{1}{L}\log(p(x_1,x_2,\ldots,x_L))}$$

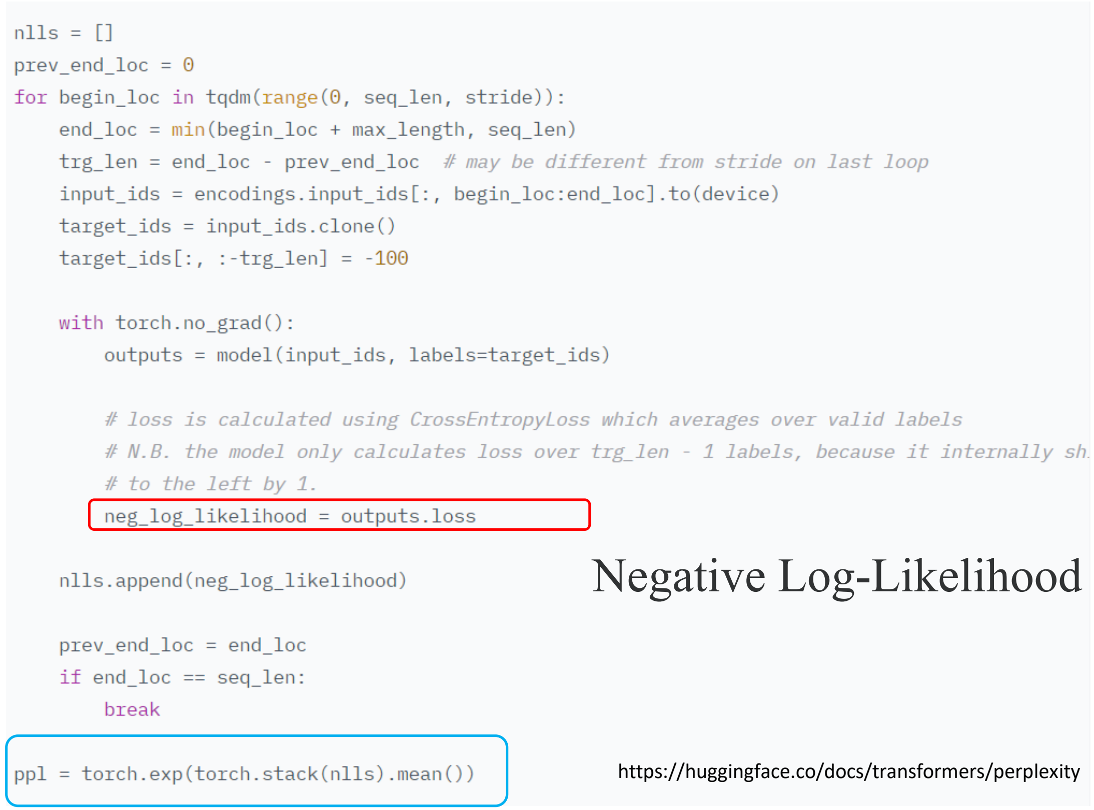
$$=exp\left\{-\frac{1}{L}\sum_{i=1}^{L}\log(p(x_i|x_1,\ldots,x_{i-1}))\right\}$$

怎麼評估(evaluation): Perplexity

Approx.

cross entropy: $H(X) = -\sum_{t=1}^{L} y_t log_2(p_t)$

$$PPL(X) = \boxed{exp}\left\{\boxed{-\frac{1}{L}\sum_{i=1}^{L}\log(p(x_i|x_1,\dots,x_{i-1}))}\right\}$$



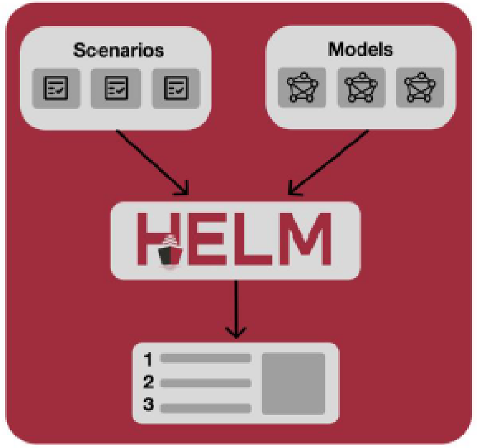
```
nlls = []
prev_end_loc = 0
for begin_loc in tqdm(range(0, seq_len, stride)):
    end_loc = min(begin_loc + max_length, seq_len)
    trg_len = end_loc - prev_end_loc  # may be different from stride on last loop
    input_ids = encodings.input_ids[:, begin_loc:end_loc].to(device)
    target_ids = input_ids.clone()
    target_ids[:, :-trg_len] = -100

    with torch.no_grad():
        outputs = model(input_ids, labels=target_ids)

        # loss is calculated using CrossEntropyLoss which averages over valid labels
        # N.B. the model only calculates loss over trg_len - 1 labels, because it internally sh
        # to the left by 1.
        neg_log_likelihood = outputs.loss

    nlls.append(neg_log_likelihood)

    prev_end_loc = end_loc
    if end_loc == seq_len:
        break

ppl = torch.exp(torch.stack(nlls).mean())
```

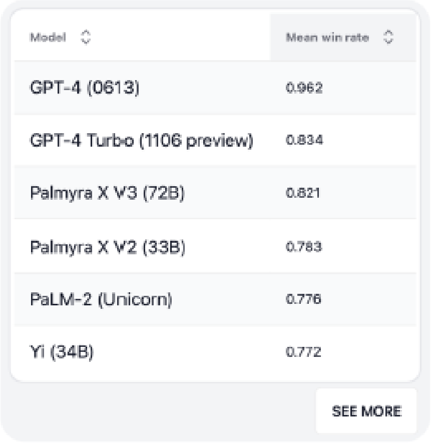
Negative Log-Likelihood

https://huggingface.co/docs/transformers/perplexity

怎麼評估LLM
Holistic evaluation of language models (HELM)(2022)

Holistic evaluation of language models (HELM)

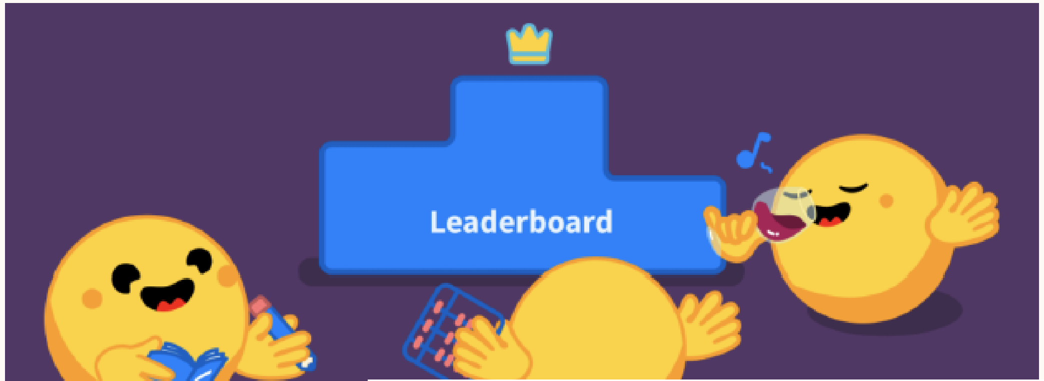


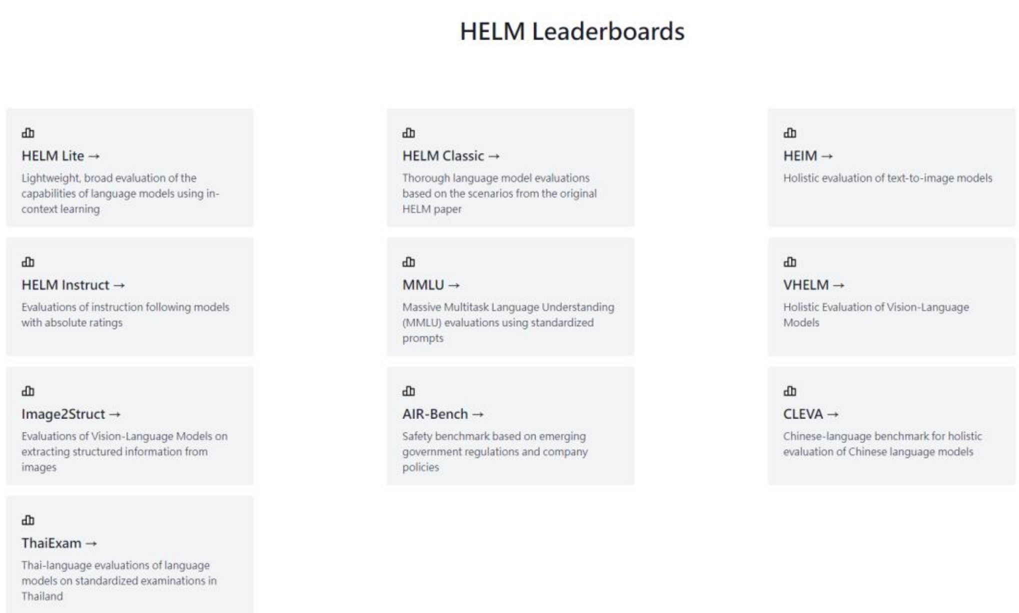


Model	Mean win rate
GPT-4 (0613)	0.962
GPT-4 Turbo (1106 preview)	0.834
Palmyra X V3 (72B)	0.821
Palmyra X V2 (33B)	0.783
PaLM-2 (Unicorn)	0.776
Yi (34B)	0.772

https://crfm.stanford.edu/helm/

Huggingface open LLM leaderboard





HELM Leaderboards

HELM Lite →
Lightweight, broad evaluation of the capabilities of language models using in-context learning

HELM Classic →
Thorough language model evaluations based on the scenarios from the original HELM paper

HEIM →
Holistic evaluation of text-to-image models

HELM Instruct →
Evaluations of instruction following models with absolute ratings

MMLU →
Massive Multitask Language Understanding (MMLU) evaluations using standardized prompts

VHELM →
Holistic Evaluation of Vision-Language Models

Image2Struct →
Evaluations of Vision-Language Models on extracting structured information from images

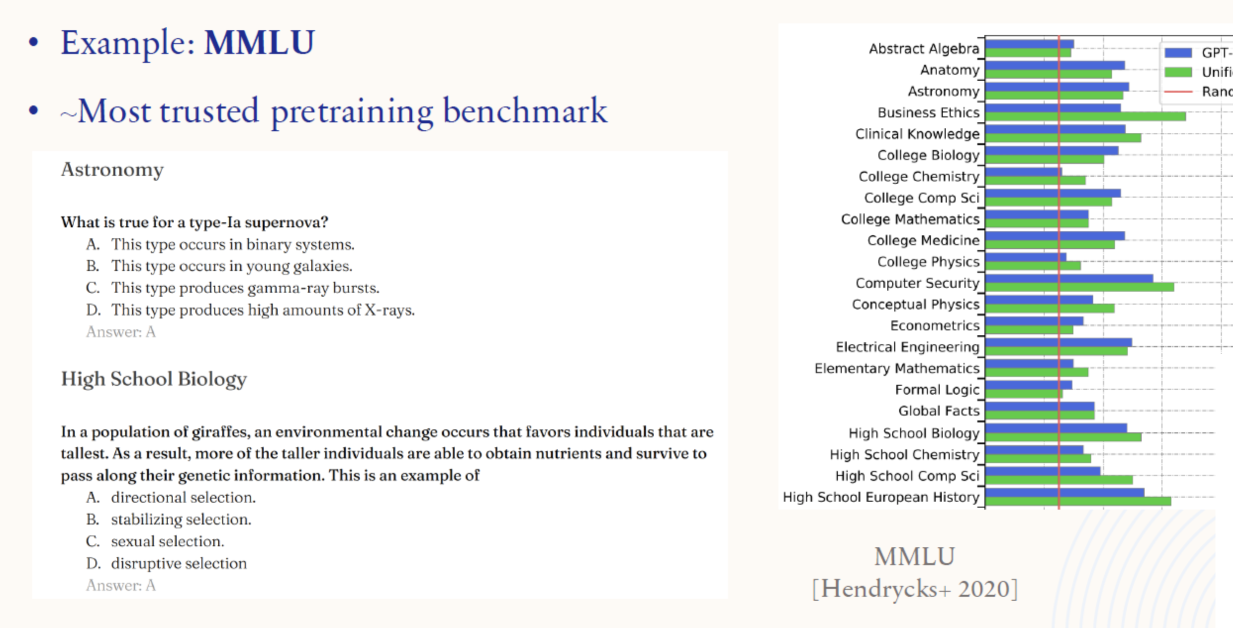
AIR-Bench →
Safety benchmark based on emerging government regulations and company policies

CLEVA →
Chinese-language benchmark for holistic evaluation of Chinese language models

ThaiExam →
Thai-language evaluations of language models on standardized examinations in Thailand

Stanford CS229 I Machine Learning I Building Large Language Models (LLMs)

怎麼評估LLM: HELM-MMLU



- Example: **MMLU**
- ~Most trusted pretraining benchmark

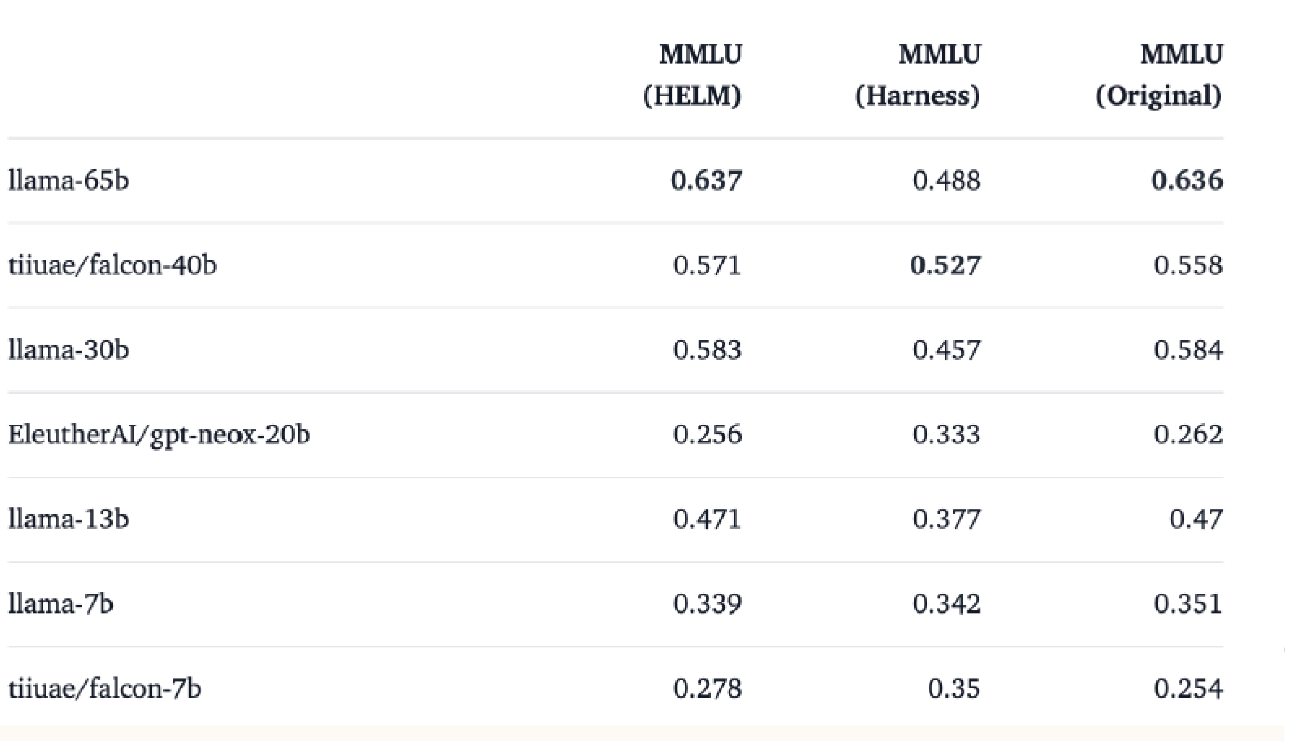
Astronomy

What is true for a type-Ia supernova?
A. This type occurs in binary systems.
B. This type occurs in young galaxies.
C. This type produces gamma-ray bursts.
D. This type produces high amounts of X-rays.
Answer: A

High School Biology

In a population of giraffes, an environmental change occurs that favors individuals that are tallest. As a result, more of the taller individuals are able to obtain nutrients and survive to pass along their genetic information. This is an example of
A. directional selection.
B. stabilizing selection.
C. sexual selection.
D. disruptive selection
Answer: A

MMLU
[Hendrycks+ 2020]



	MMLU (HELM)	MMLU (Harness)	MMLU (Original)
llama-65b	**0.637**	0.488	**0.636**
tiiuae/falcon-40b	0.571	**0.527**	0.558
llama-30b	0.583	0.457	0.584
EleutherAI/gpt-neox-20b	0.256	0.333	0.262
llama-13b	0.471	0.377	0.47
llama-7b	0.339	0.342	0.351
tiiuae/falcon-7b	0.278	0.35	0.254

Q&A