智慧運算學院
College of Intelligent Computing

人工智慧學系
Department of Artificial Intelligence

長庚大學
CGU
1987
CHANG GUNG UNIVERSITY

智慧運算技術導論

自然語言處理篇 – 大型語言模型

林英嘉 (Ying-Jia Lin)

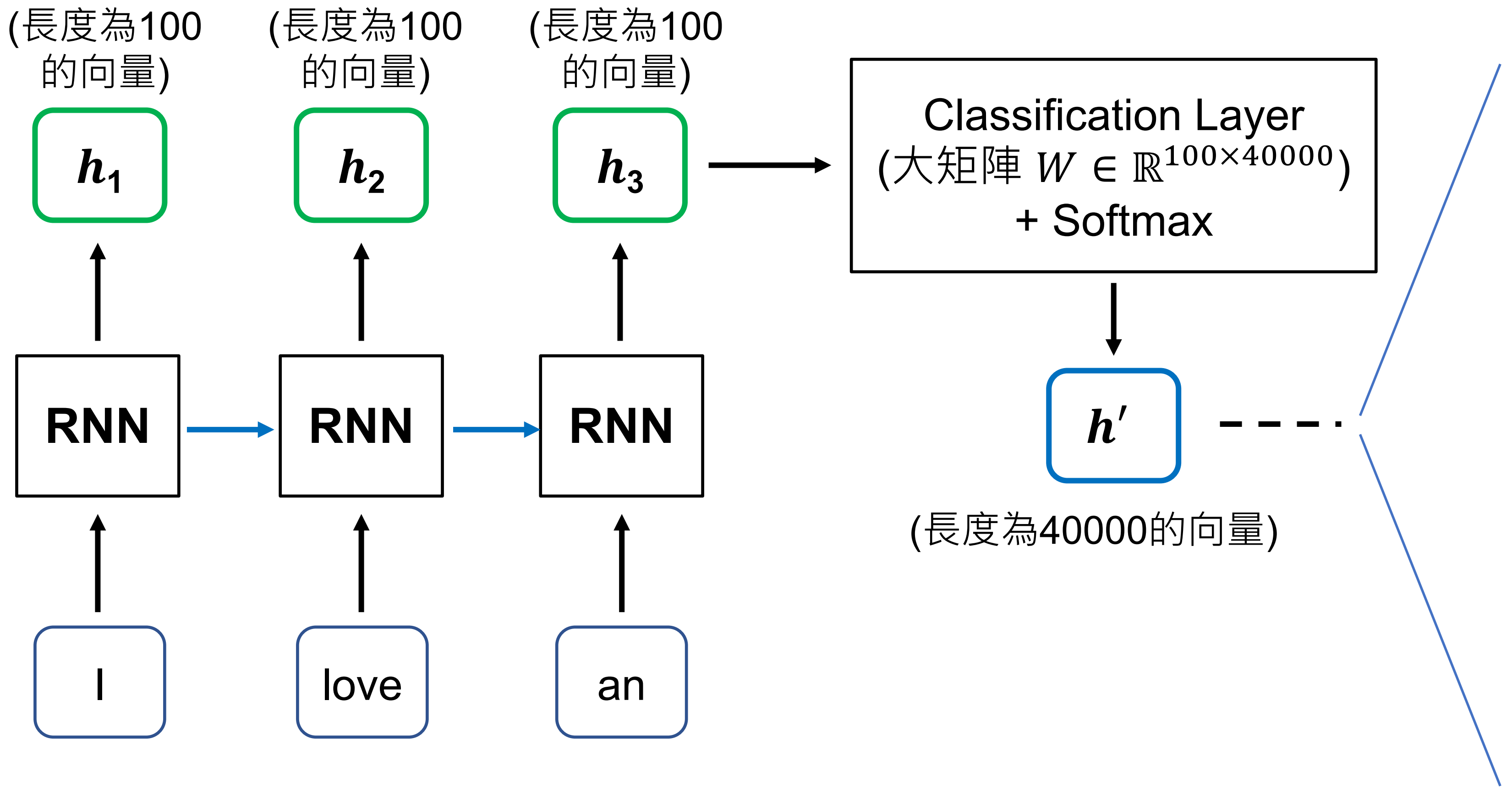
長庚大學人工智慧學系

2026/01/28

Slido
# AIMD (Link)

[Recap] 文字生成過程

假設 Hidden size 為 100；字典大小（vocab_size）為 40000



字典裡面的字	機率值
a	0.0001
…	…
apple	0.600
…	…
b	0.0001
…	…
banana	0.0003
…	…
elephant	0.300
…	…
eraser	0.05
…	…
…	…
zoo	0.0001

Vocab size

Outline

- Transformer 架構的後續應用2 – GPT
- Decoding strategies
- LLM (以 InstructGPT 為例)
- RAG

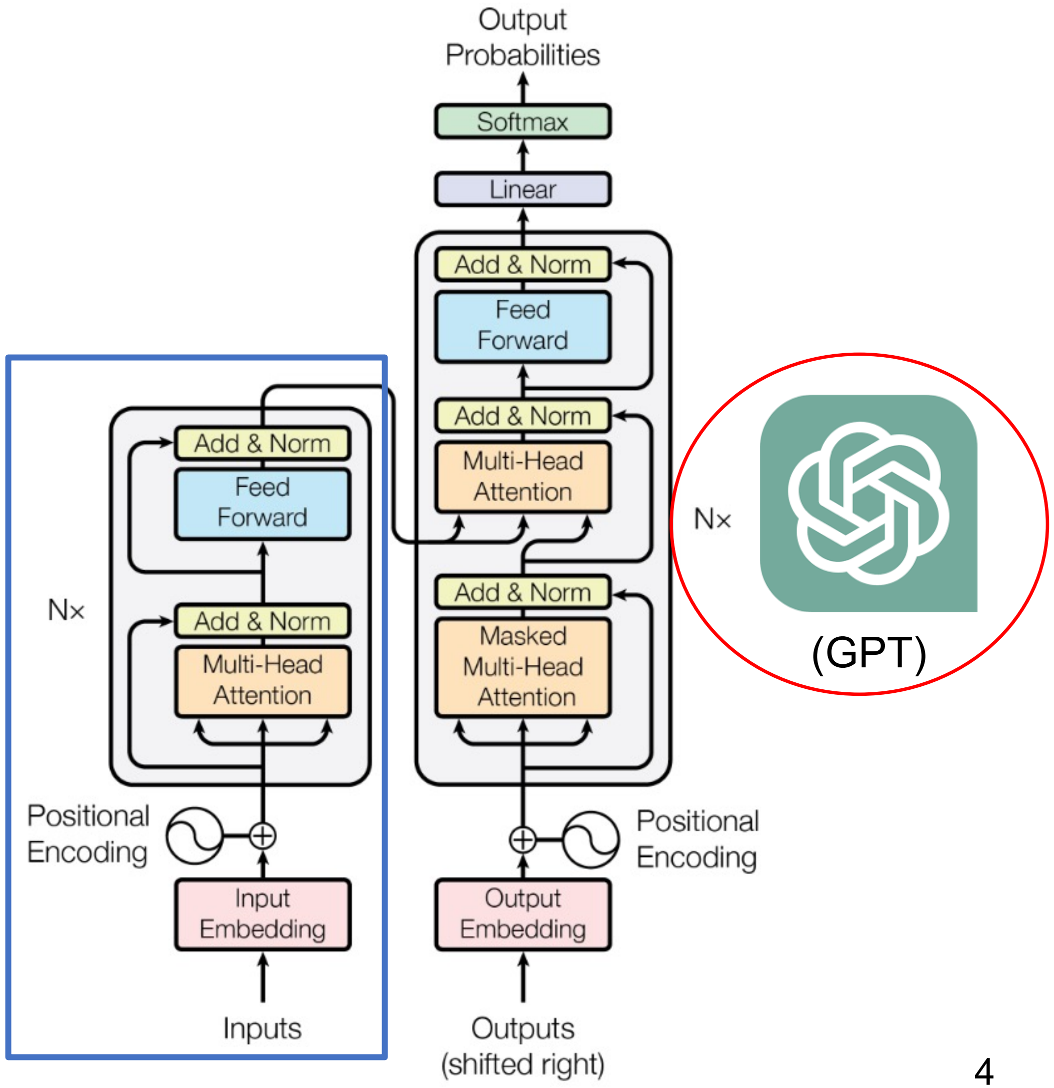
Transformer 架構的後續應用

- Generative Pre-training (GPT) series 是 **Transformer Decoder**
- BERT (Bidirectional encoder representations from transformers) 是 **Transformer Encoder**
    - [Devlin et al., 2018]

(BERT)

N=12

Encoder

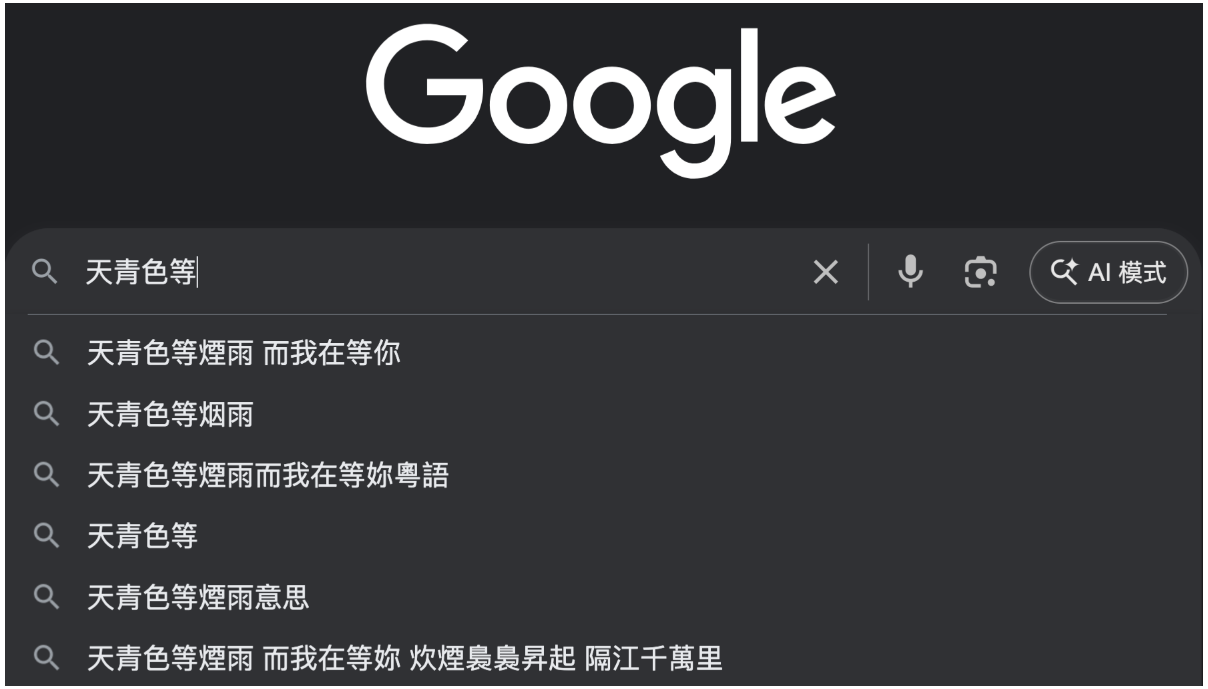


GPT 是一個語言模型

https://developers.google.com/machine-learning/resources/intro-llms

- 語言模型定義：語言模型 (Language Model) 是一個能預測並生成語言的機器學習模型

(例子) 下一個字預測 ➡️



天青色等

天青色等煙雨 而我在等你

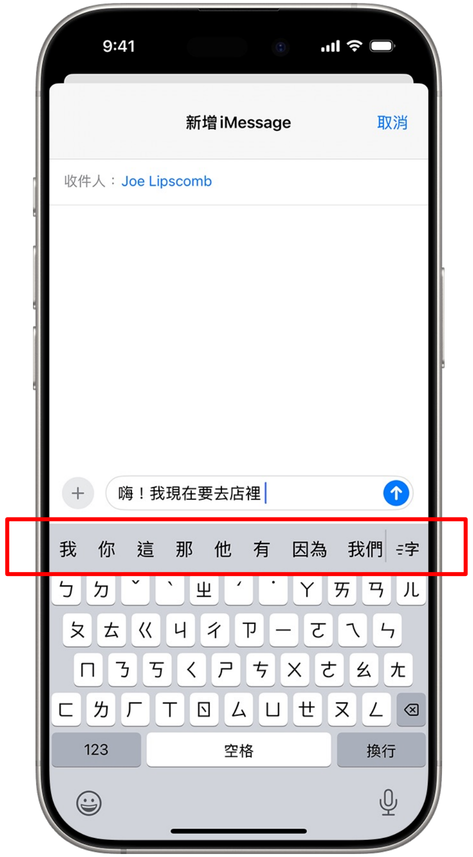
天青色等烟雨

天青色等煙雨而我在等妳粵語

天青色等

天青色等煙雨意思

天青色等煙雨 而我在等妳 炊煙裊裊昇起 隔江千萬里



收件人：Joe Lipscomb

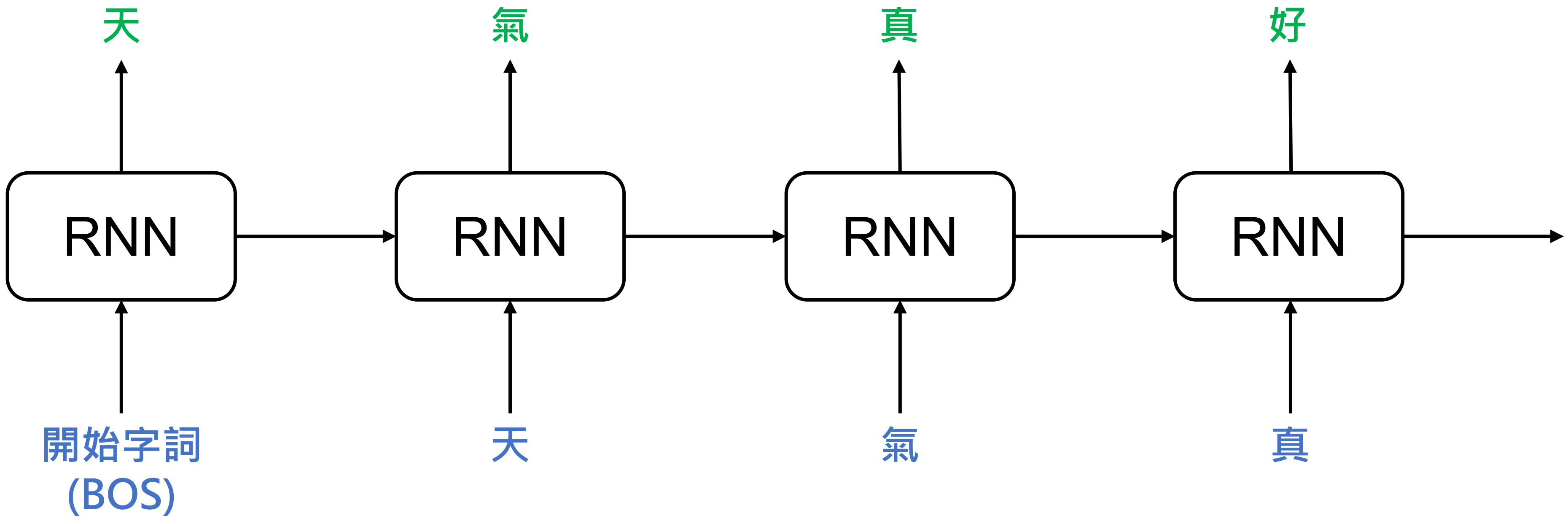
嗨！我現在要去店裡

我 你 這 那 他 有 因為 我們

Figure source: https://support.apple.com/zh-tw/104995

[RNN Recap] 條件機率

假設要讓模型生成一個句子：「天氣真好」



P(天|BOS)　　*P*(氣|BOS、天)　　*P*(真|BOS、天、氣)　　*P*(好|BOS、天、氣、真)

NLP

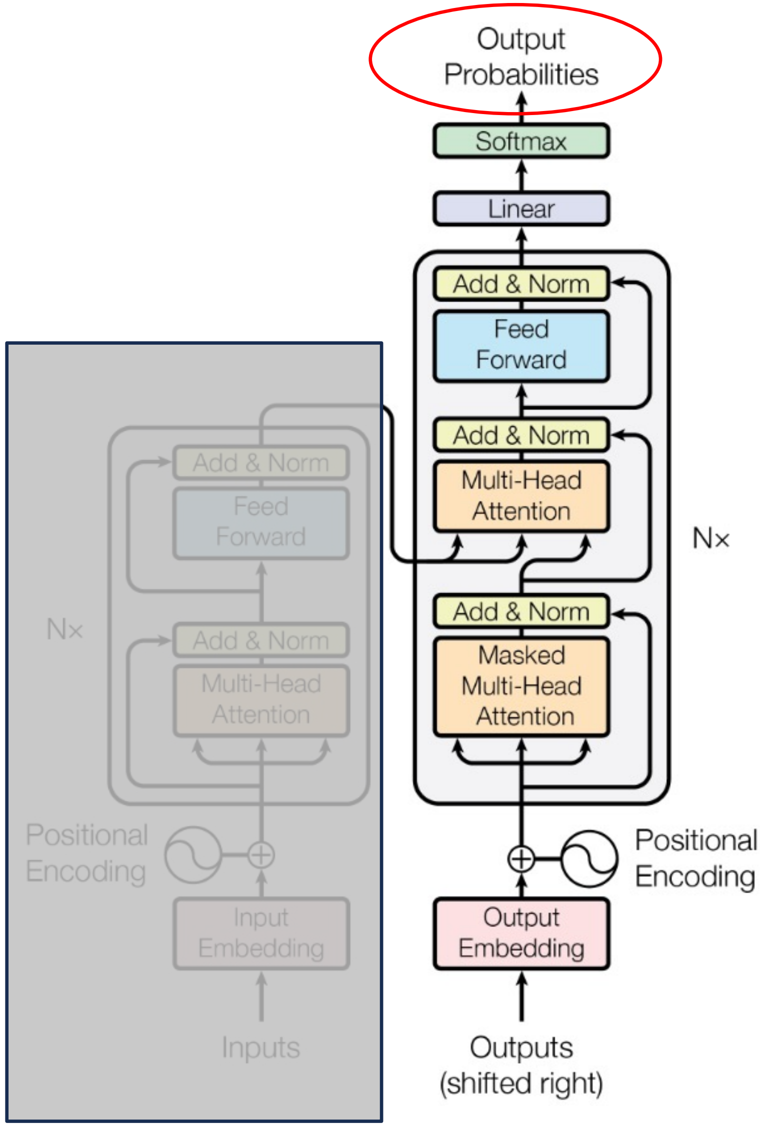
常見 Language Models

💡**Language Models: 透過 Language Modeling 去預測詞彙機率分佈的模型**

RNN

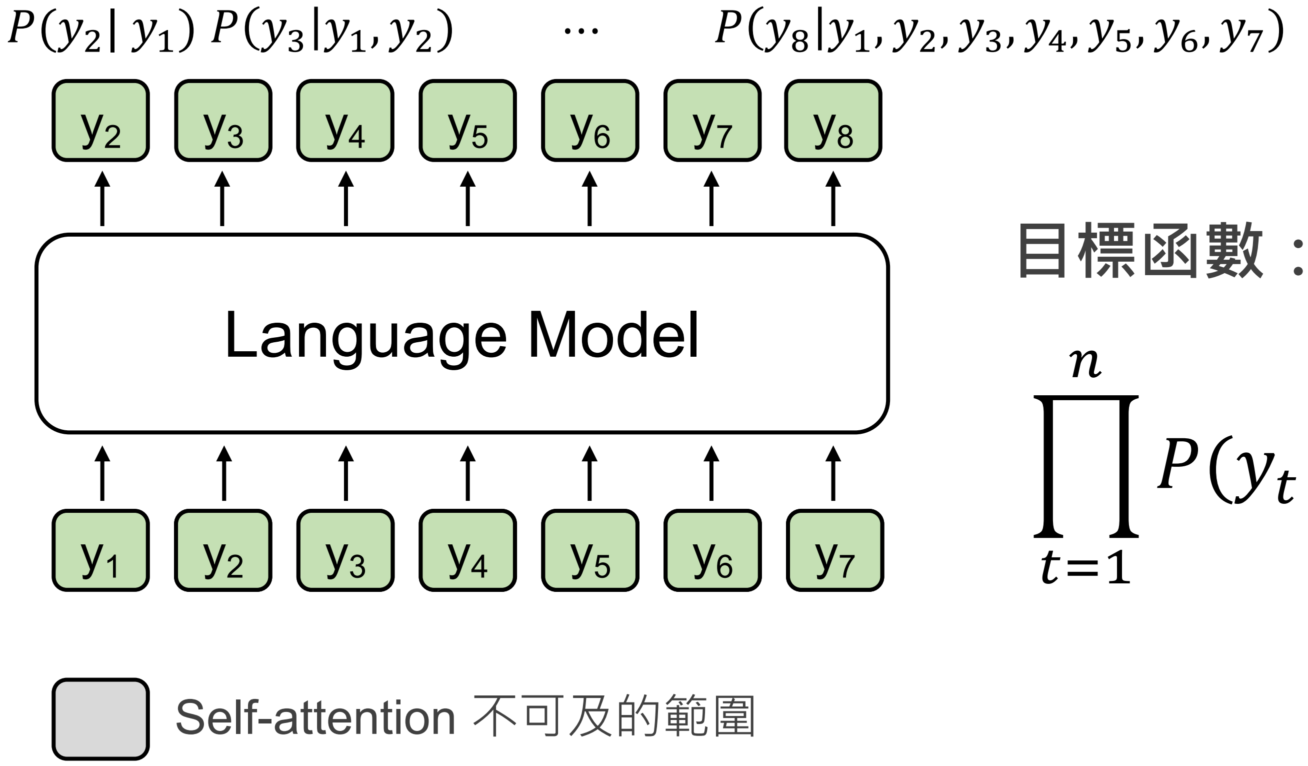
Transformer

GPT-series



NLP

訓練 GPT 最大化每個時間點的機率


$P(y_2|\, y_1)$ $P(y_3|y_1, y_2)$ … $P(y_8|y_1, y_2, y_3, y_4, y_5, y_6, y_7)$
y2 y3 y4 y5 y6 y7 y8
Language Model
y1 y2 y3 y4 y5 y6 y7
Self-attention 不可及的範圍

目標函數：

$$\prod_{t=1}^{n} P(y_t|y_1, y_2, \dots, y_{t-1}) \longleftarrow \text{Language Modeling}$$

Language Modeling and Cross-entropy (1)

為了使語言模型能夠以分類的形式被訓練，**通常會取log**

$$\log(\prod_{t=1}^{n} P(y_t|y_1, y_2, \ldots, y_{t-1}))$$

$$= \sum_{t=1}^{n} \log P(y_t|y_1, y_2, \ldots, y_{t-1}) \longleftarrow$$

越大越好

加上負號之後就是 **Cross-entropy**

(loss 越小越好)

Language Modeling and Cross-entropy (2)

Language modeling 在 t 時間點的正確答案是 t+1 的字詞

I love CGU

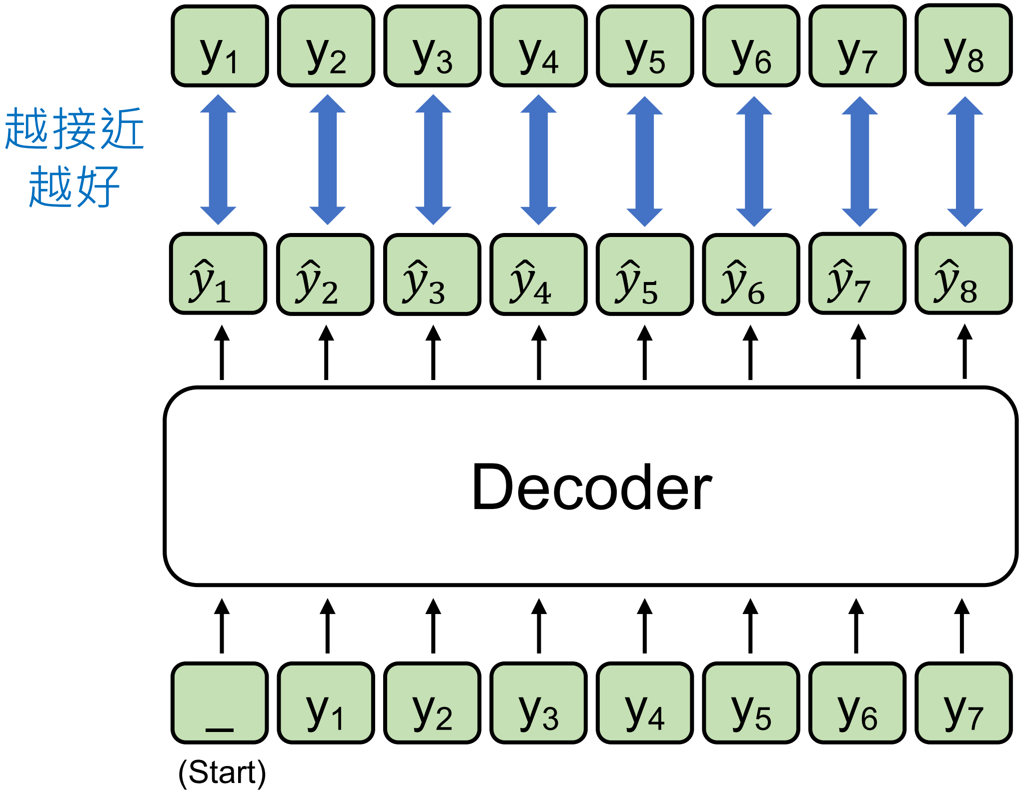
t t+1

Example Vocab (假設只有5個字)

Token	Logits	取 Softmax 後之機率	正確答案
CGU	0.0011	0.78	1
I	0.0012	0.11	0
love	0.0013	0.03	0
hate	0.0014	0.02	0
like	0.0015	0.06	0

➡ Cross-entropy

Language Models 終究是 left-to-right



越接近
越好
y_1 y_2 y_3 y_4 y_5 y_6 y_7 y_8
$\hat{y}_1$ $\hat{y}_2$ $\hat{y}_3$ $\hat{y}_4$ $\hat{y}_5$ $\hat{y}_6$ $\hat{y}_7$ $\hat{y}_8$
Decoder
_ y_1 y_2 y_3 y_4 y_5 y_6 y_7
(Start)

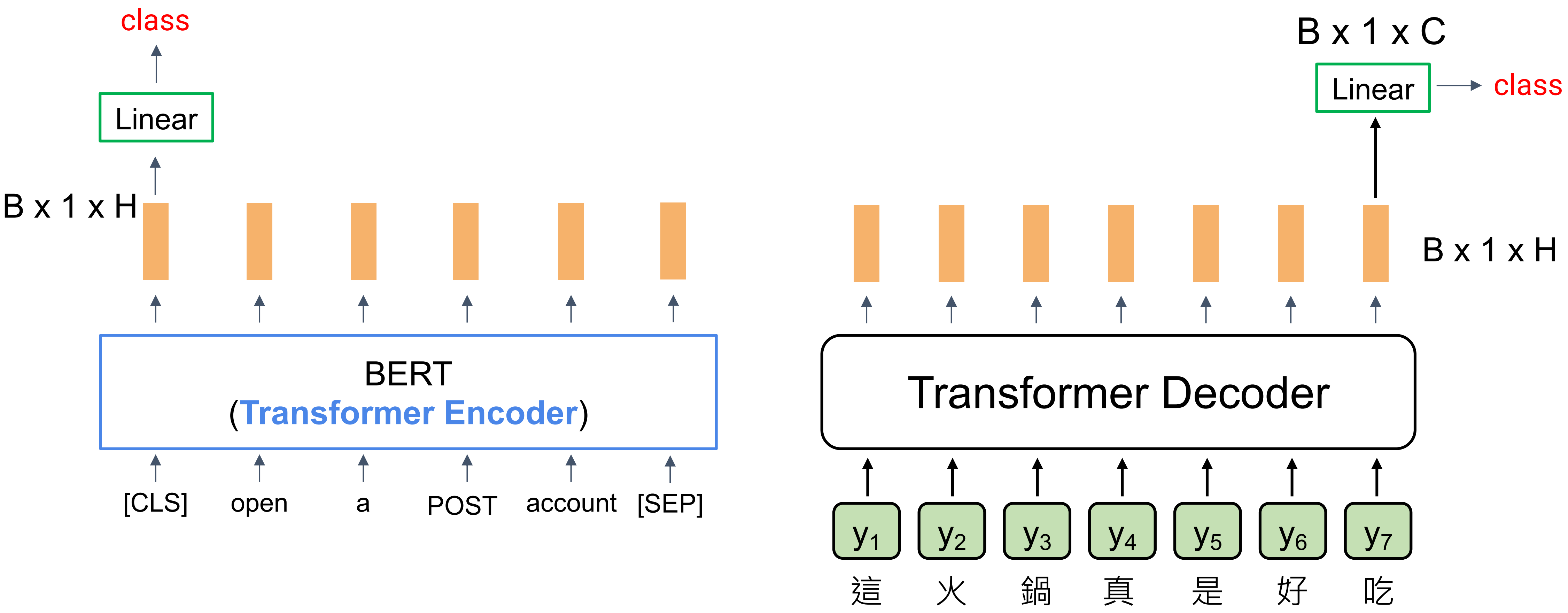
(2018年的Google)
GPT-1 不好，因為訓練無雙向，BERT 最棒

Encoder vs. Decoder in Pre-training

	Encoder (e.g., BERT)	Decoder (e.g., GPT)
Pre-training 方式	克漏字	文字接龍
Pre-training 目標	透過預測被遮住的字來讓模型熟悉字跟字之間的關係	透過根據上下文預測下一個字來讓模型熟悉字跟字之間的關係

GPT-1 與 BERT 的 Fine-tuning

以 Classification 為例

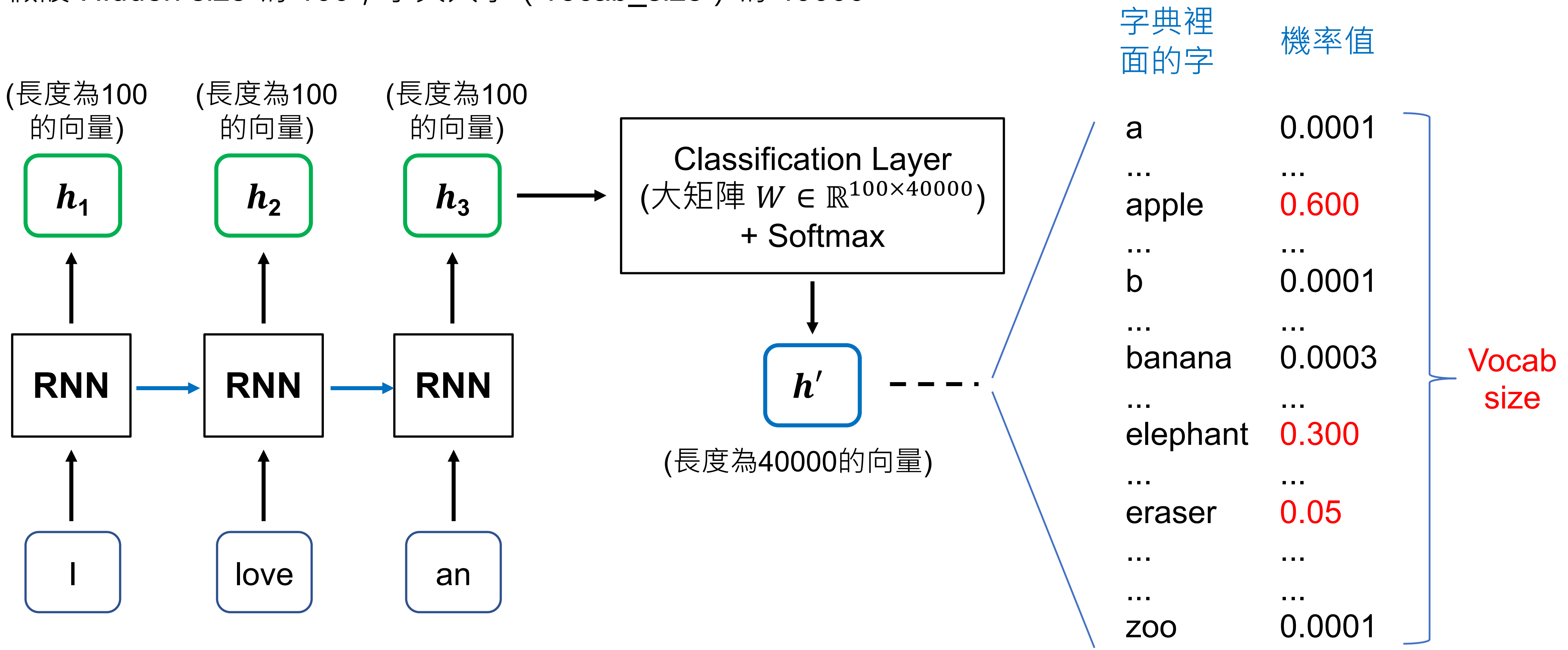


Decoding Strategies

- Greedy Decoding
- Beam Search
- Top-k Sampling
- Top-p Sampling

文字生成過程 – Greedy Decoding

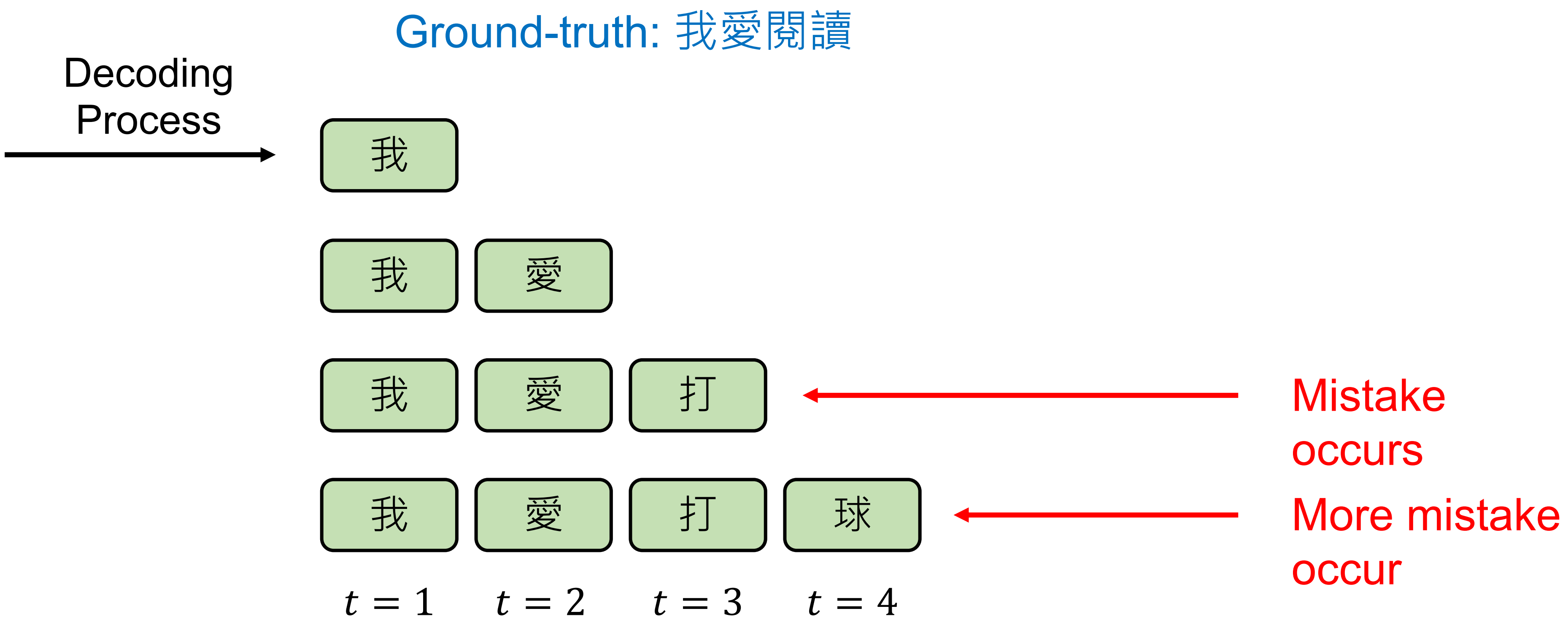
假設 Hidden size 為 100；字典大小（vocab_size）為 40000



字典裡面的字	機率值
a	0.0001
…	…
apple	0.600
…	…
b	0.0001
…	…
banana	0.0003
…	…
elephant	0.300
…	…
eraser	0.05
…	…
…	…
zoo	0.0001

Problem of Greedy Decoding

- Greedy decoding cannot undo!



Ground-truth: 我愛閱讀

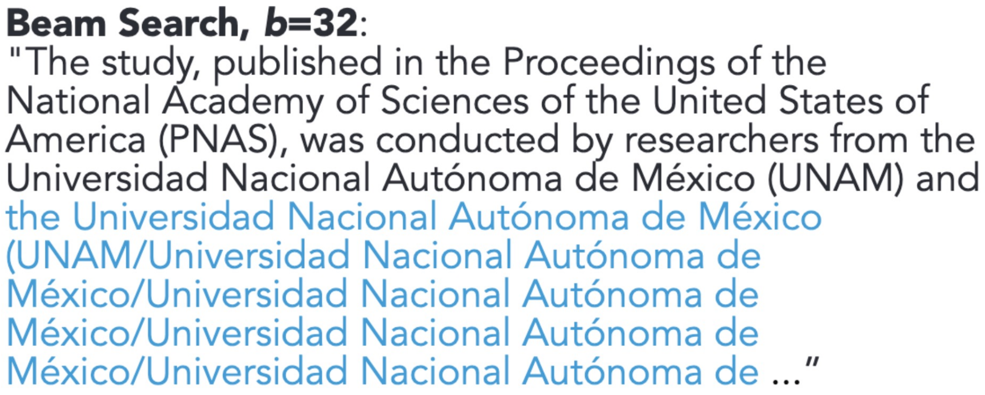
Decoding Process

Mistake occurs

More mistake occur

Problem of Beam Search (Greedy Decoding的改良版)

Context: In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.



Beam Search, *b*=32:
"The study, published in the Proceedings of the National Academy of Sciences of the United States of America (PNAS), was conducted by researchers from the Universidad Nacional Autónoma de México (UNAM) and the Universidad Nacional Autónoma de México (UNAM/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de ..."

Holtzman, Ari, et al. "The Curious Case of Neural Text Degeneration." International Conference on Learning Representations (ICLR). 2020.

思考 Greedy Decoding 的問題

- 現代語言模型通常使用 maximum likelihood 的方式 (language modeling) 進行訓練，這會導致模型過度偏向常見或高頻 tokens
- 當模型在 early steps 中對某些 tokens 給予極高機率時，這些 tokens 所在的路徑就會大幅壓制了其他 candidate tokens，導致生成缺少多樣性，甚至進入重複生成的loop (例如 I don’t know I don’t know I don’t know …)

Holtzman, Ari, et al. "The Curious Case of Neural Text Degeneration." International Conference on Learning Representations (ICLR). 2020.

[Summary] Strategy for MLE decoding

Maximum Likelihood Estimation (MLE): greedy decoding, beam search

既然機率值最高的不是我們想要的字詞

Strategy: Add more randomness!

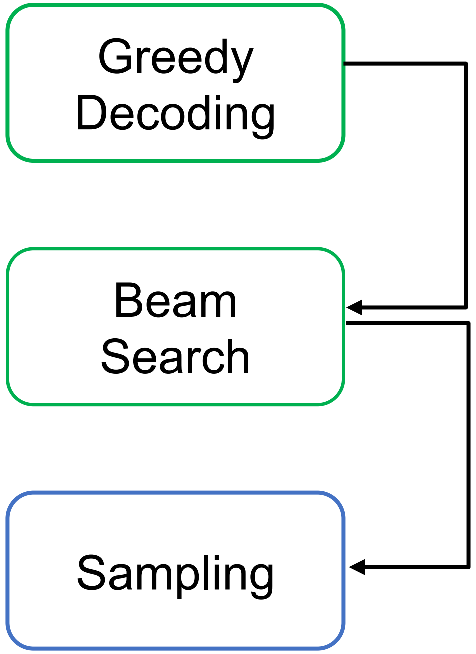
Strategy for MLE decoding: 改用 Sampling

- Sampling 就是「讓模型隨機選下一個詞，而不是每次都選最有可能的那個詞」。
- Sampling 是根據模型對每個詞給出的 機率分布 來進行「隨機抽樣」

Token	Probability (p)
cat	0.5
dog	0.3
car	0.15
book	0.05

Sampling 不是亂抽！
而是根據機率 (模型的信心) 來產生下一次生成

Summary and the Thinking Route



Greedy decoding always selects the best choice at each time step and becomes problematic for early mistakes.

Models can fall into repetitive loops with beam search, especially for longer text.

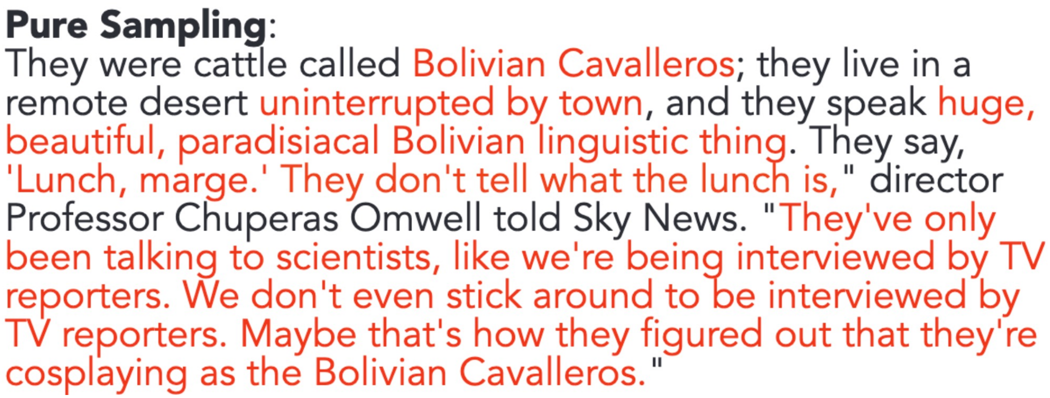
Problem of Pure Sampling

獨角獸

Context: In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English. 英語

山谷

如果完全 random sampling



Pure Sampling:
They were cattle called Bolivian Cavalleros; they live in a remote desert uninterrupted by town, and they speak huge, beautiful, paradisiacal Bolivian linguistic thing. They say, 'Lunch, marge.' They don't tell what the lunch is," director Professor Chuperas Omwell told Sky News. "They've only been talking to scientists, like we're being interviewed by TV reporters. We don't even stick around to be interviewed by TV reporters. Maybe that's how they figured out that they're cosplaying as the Bolivian Cavalleros."

它們是被稱為「玻利維亞騎士」的牛；他們住在遠離城鎮的偏遠沙漠，他們說話很大聲，美麗、天堂般的玻利維亞語言。他們說，「吃午餐了，瑪姬。」他們沒有告訴午餐是什麼，」導演丘佩拉斯·奧姆威爾教授告訴天空新聞。「他們只是和科學家交談，就像接受電視採訪一樣記者。我們甚至沒有留下來接受採訪電視記者。也許這就是他們發現他們扮演玻利維亞騎士。」

Holtzman, Ari, et al. "The Curious Case of Neural Text Degeneration." International Conference on Learning Representations (ICLR). 2020.

Why is Pure Sampling so weak?

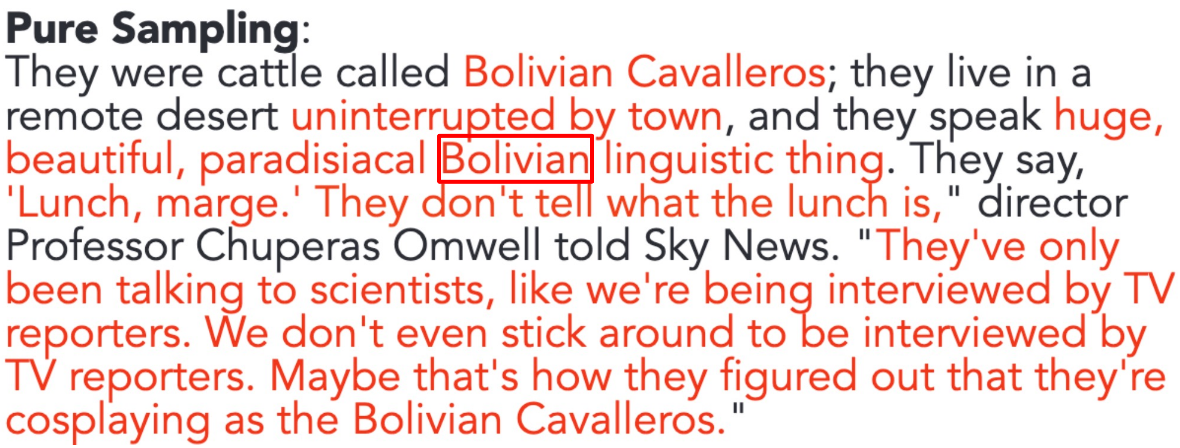
- Pure Sampling does not show repetitive loop, but the result becomes incoherent and almost unrelated to the context
- Why? -> Unreliable tail

Words that have low probabilities

Token	Prob
the	0.0011
am	0.0012
no	0.0013
a	0.0014
/	0.0015
...	...
...	...
...	...
...	...
...	...
...	...
...	...
...	...
...	...

Example Vocab

[Summary] Strategy for Sampling



Pure Sampling:
They were cattle called Bolivian Cavalleros; they live in a remote desert uninterrupted by town, and they speak huge, beautiful, paradisiacal Bolivian linguistic thing. They say, 'Lunch, marge.' They don't tell what the lunch is," director Professor Chuperas Omwell told Sky News. "They've only been talking to scientists, like we're being interviewed by TV reporters. We don't even stick around to be interviewed by TV reporters. Maybe that's how they figured out that they're cosplaying as the Bolivian Cavalleros."

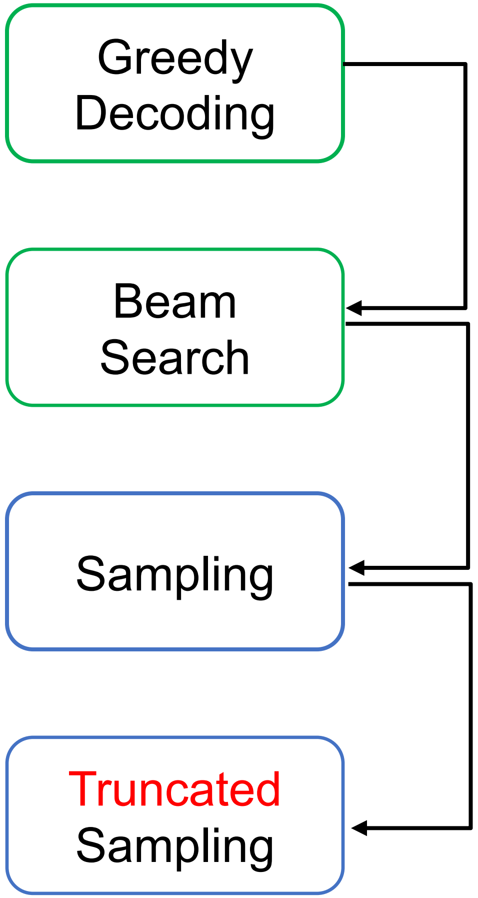
Token	Prob
Bolivian	0.0011
Chinese	0.0012
Egyptian	0.0013
...	...
...	...
...	...
...	...
...	...
...	...
...	...
...	...
...	...
Scottish	0.09
English	0.10

Bad candidates
可能有很多字
但機率總和才10%

Good candidates
可能沒有很多字
但機率總和有90%

Let's truncate the vocabulary!

Summary and the Thinking Route



Greedy Decoding

Beam Search

Sampling

Truncated Sampling

Greedy decoding always selects the best choice at each time step and becomes problematic for early mistakes.

Models can fall into repetitive loops with beam search, especially for longer text.

Random sampling takes words with low probabilities into considerations.

Top-p Sampling

Holtzman, Ari, et al. "The Curious Case of Neural Text Degeneration." International Conference on Learning Representations. 2020.

- 又稱作 Nucleus Sampling
- Core idea: truncate the vocabulary based on probability mass
- Steps:
  1. Define a value p as the probability threshold.
  2. Truncate the vocabulary whose sum of probabilities is greater than *p* :

$$\sum_{x \in V^{(p)}} P(x | x_{1:x-1}) \geq p$$

where $V^{(p)} \subset V$ is the smallest set that fulfills the equation.
Now you get a new truncated vocabulary $V^{(p)}$.

Top-p Sampling

- Core idea: truncate the vocabulary based on probability mass
- Steps:

3. Re-build the probability distribution based on the following normalization:

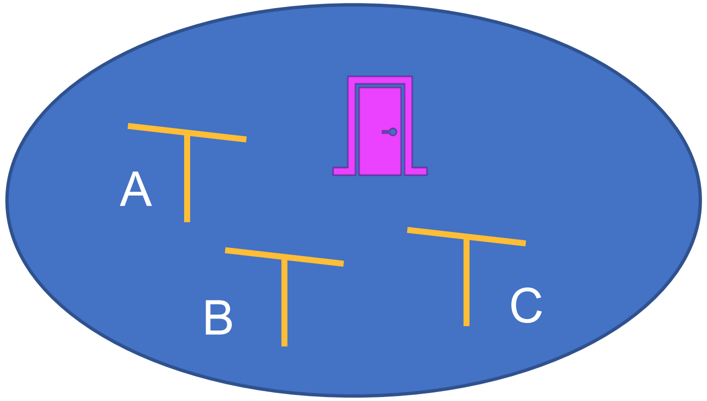
3-1. 把 $V^{(p)}$ 的機率值加總

3-2. $V^{(p)}$ 內的每個機率值/加總

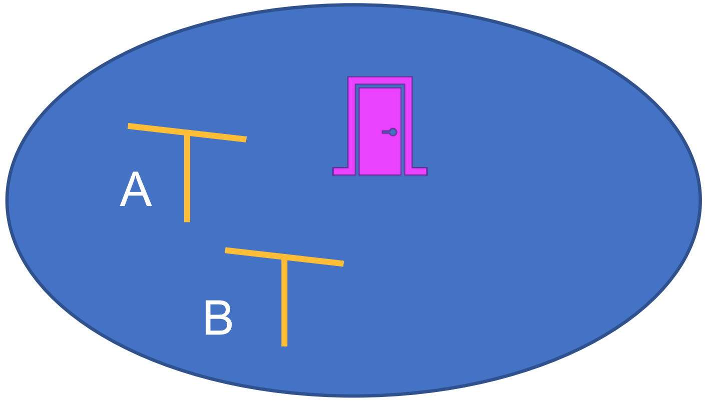
$$p' = \sum_{x \in V^{(p)}} P(x|x_{1:x-1})$$

$$\mathrm{P}'(\mathrm{x}|x_{1:x-1}) = \begin{cases} P(x|x_{1:x-1}/p') & \text{if } x \in V^{(p)} \\ 0 & \text{otherwise} \end{cases}$$

機率重新調整 toy example



任意門的機率：1/4



任意門的機率：0.25 / (0.25+0.25+0.25) = 1/3

竹蜻蜓A機率：0.25 / (0.25+0.25+0.25) = 1/3
竹蜻蜓 B 機率：0.25 / (0.25+0.25+0.25) = 1/3

Top-p Sampling example

Token	Probability (p)
cat	0.5
dog	0.3
car	0.15
book	0.05

- 設 p 為 0.8
  - 那 $V^{(p)}$ 中只會有 cat 和 dog 兩個詞

Top-k Sampling

- Core idea: truncate the vocabulary with the most probable words
- Steps:

1. Define a value k as the size of truncated vocabulary.
2. Leave the k words with the highest probabilities. Now you get a new truncated vocabulary $V^{(k)}$. (假設k=40，那$V^{(k)}$就只剩40個 tokens)
3. Re-build the probability distribution based on the following normalization:

3-1. 把 $V^{(k)}$ 的機率值加總

3-2. $V^{(k)}$ 內的每個機率值/加總

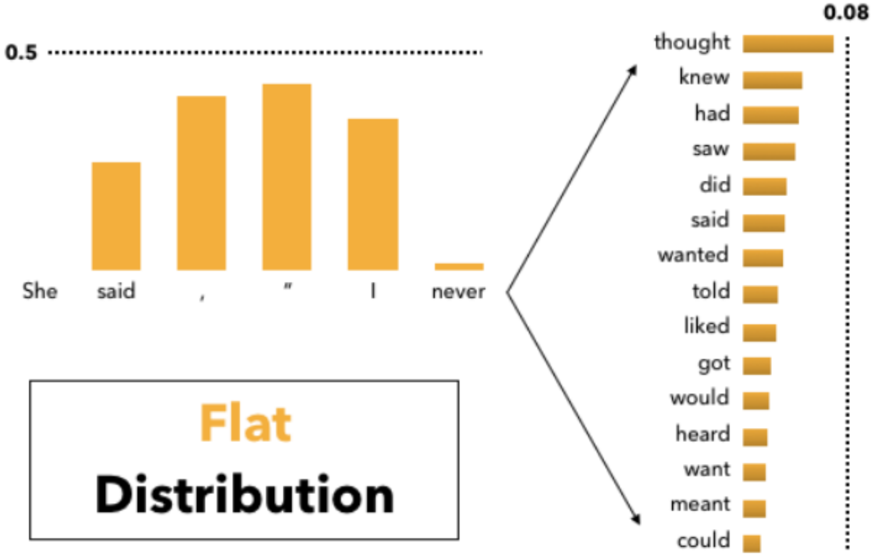
$$p' = \sum_{x \in V^{(k)}} P(x|x_{1:x-1})$$

$$\mathrm{P}'(\mathrm{x}|x_{1:x-1}) = \begin{cases} P(x|x_{1:x-1})/p' & \text{if } x \in V^{(k)} \\ 0 & \text{otherwise} \end{cases}$$

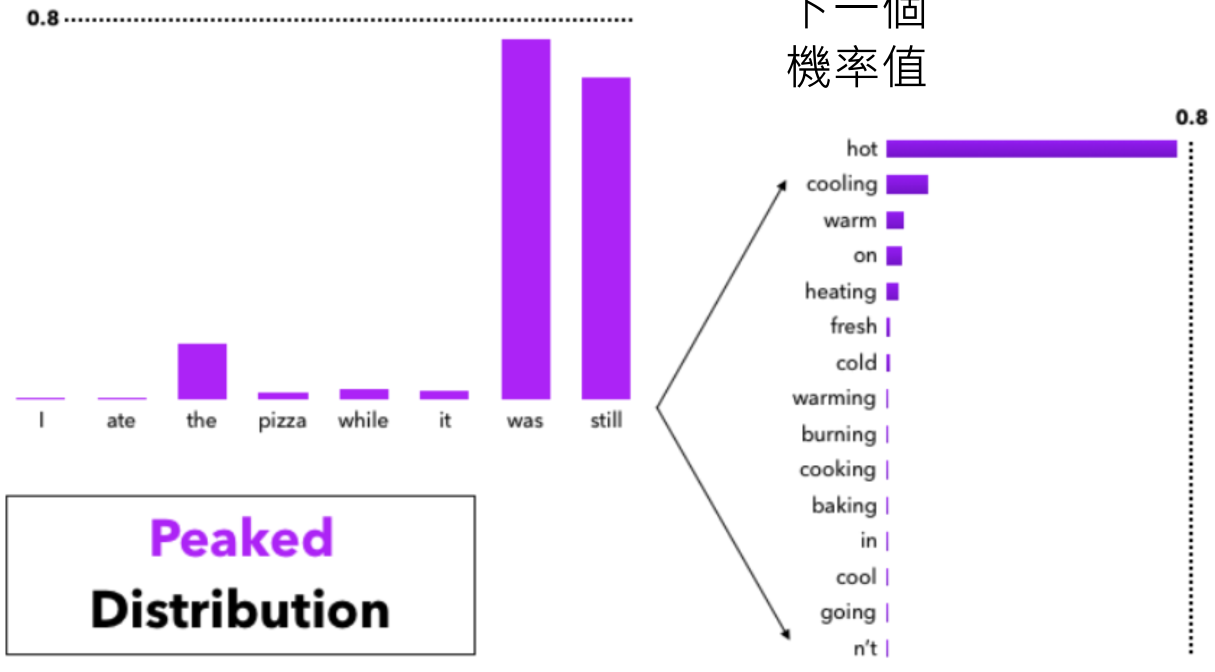
Problem of Top-k Sampling

- Top-k sampling 的 k 值需根據情況調整

下一個
機率值



下一個
機率值

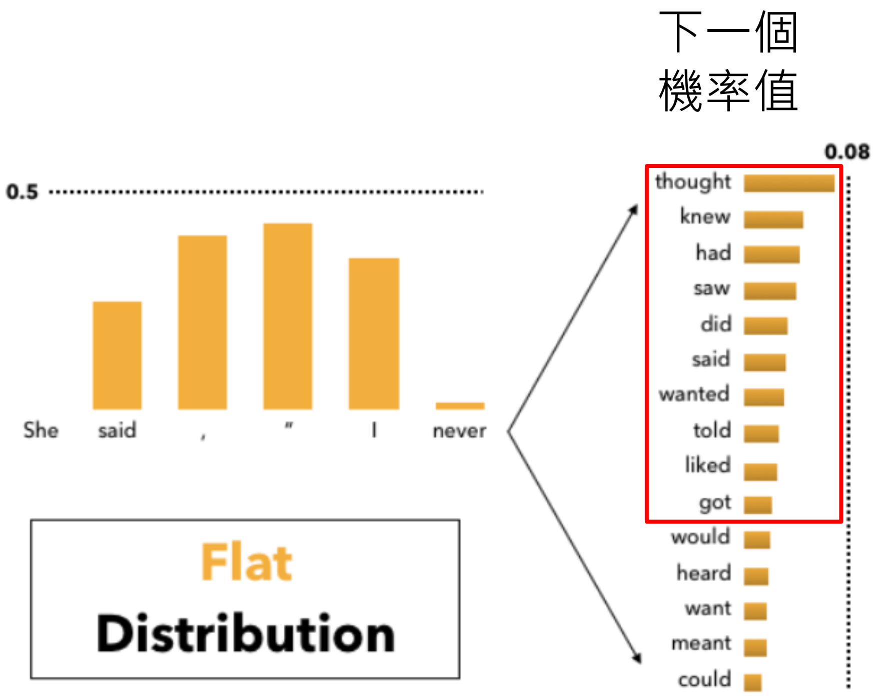


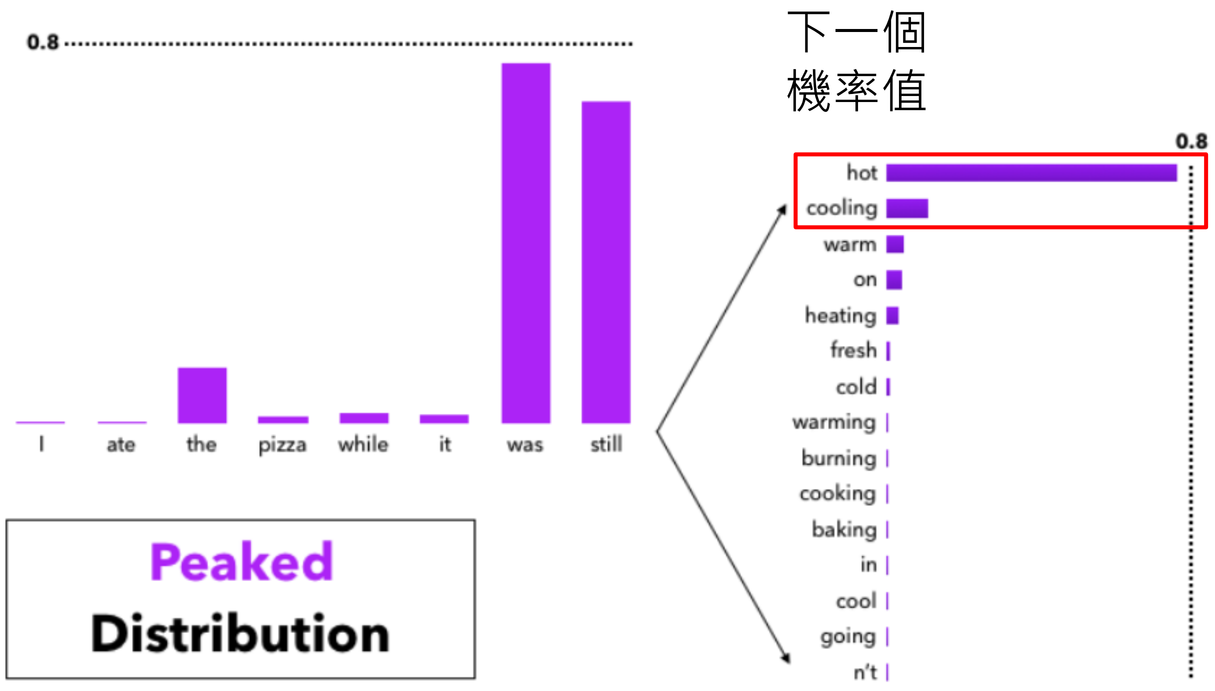
此時 k 如果太小的話，容易錯過最正確的字 -> k 要設大一點

此時 k 如果太大的話，容易取到與上下文無關的字 -> k 要設小一點

Problem of Top-k Sampling (看看 Top-p)

- 假設 p = 0.9





此時 k 如果太小的話，容易錯過最正確的字 -> k 要設大一點

此時 k 如果太大的話，容易取到與上下文無關的字 -> k 要設小一點

NLP

Softmax

- When generating the next word, `softmax` is performed to get the probabilities among the words in the vocabulary.
- Softmax formula:

$$\frac{\exp(u_l/\mathrm{t})}{\sum_{l'}^{|V|} \exp\,(u_{l'}/t)}$$

u_l: logits (model outputs before softmax)
$|V|$: size of the vocabulary
t: softmax temperature

Softmax

Exampl e Vocab	Token	Logits		Probability
	the	0.0011	→	0.78
	am	0.0012	→	0.11
	no	0.0013	→	0.03
	a	0.0014	→	0.02
	/	0.0015	→	0.06

- Note that softmax is required for every decoding strategies since we need to find out the next word from a vocabulary.

Properties of Softmax Temperature

u_l: logits (model outputs before softmax)
$|V|$: size of the vocabulary
t: softmax temperature

- Softmax formula:

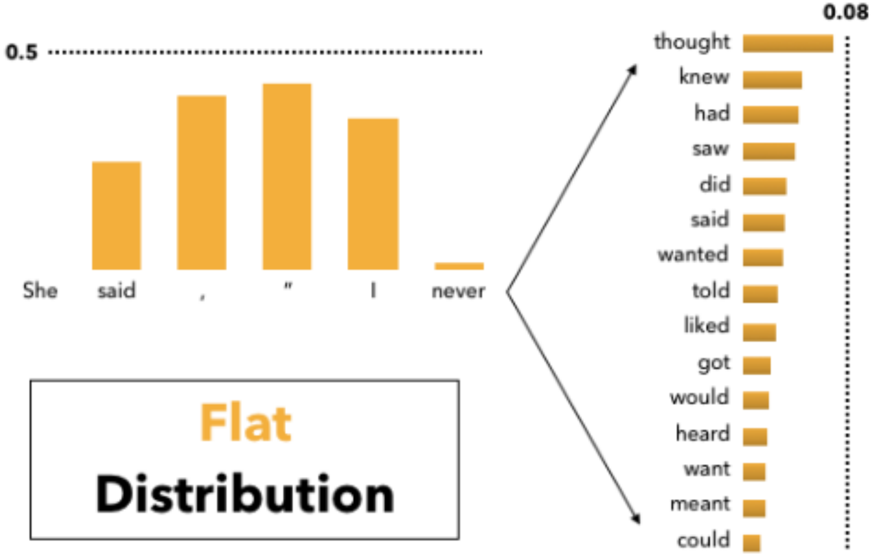
$$\frac{\exp(u_l/\mathrm{t})}{\sum_{l'}^{|V|} \exp\,(u_{l'}/t)}$$

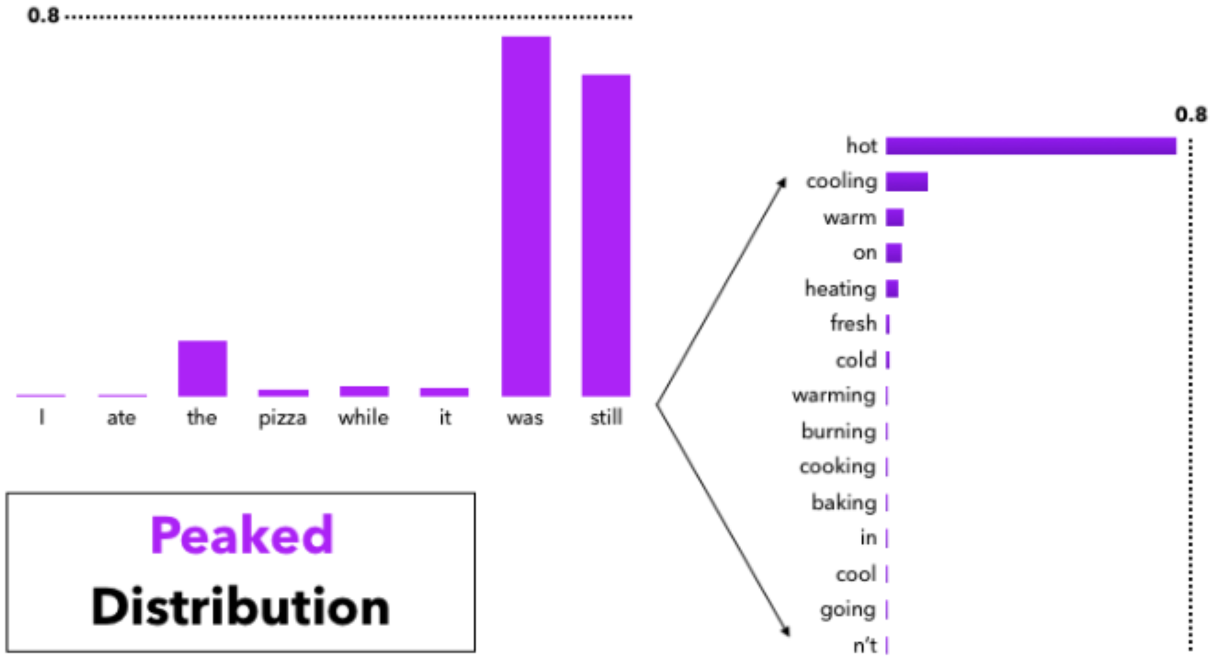
- Larger t -> Lower probability value -> Smaller ranges of probability distribution -> More diverse Outputs
- Smaller t -> Higher probability value -> Greater ranges of probability distribution -> Less diverse Outputs

Softmax Temperature

Temperature 高的時候 (more diverse)

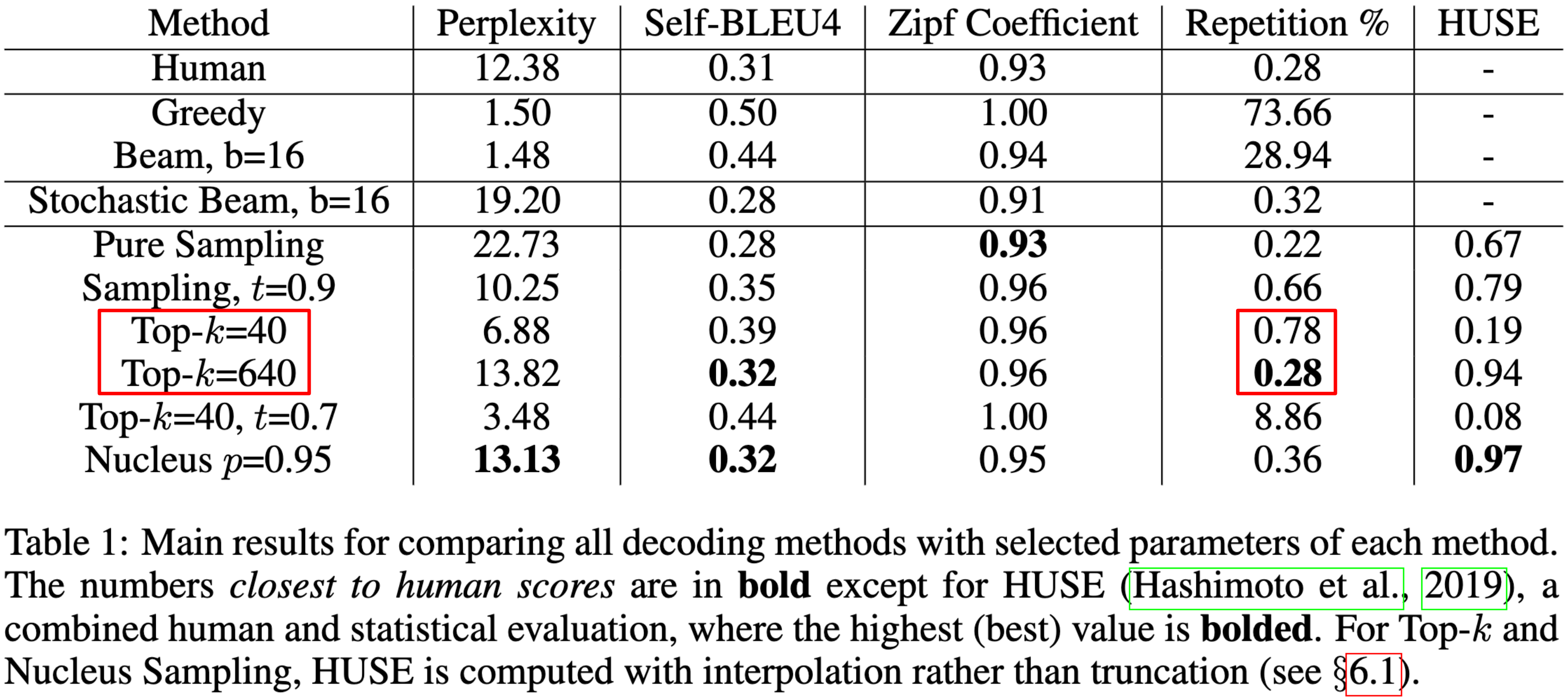
Temperature 低的時候 (less diverse)





Holtzman, Ari, et al. "The Curious Case of Neural Text Degeneration." International Conference on Learning Representations (ICLR). 2020.

The k value should be carefully picked



Method	Perplexity	Self-BLEU4	Zipf Coefficient	Repetition %	HUSE
Human	12.38	0.31	0.93	0.28	-
Greedy	1.50	0.50	1.00	73.66	-
Beam, b=16	1.48	0.44	0.94	28.94	-
Stochastic Beam, b=16	19.20	0.28	0.91	0.32	-
Pure Sampling	22.73	0.28	**0.93**	0.22	0.67
Sampling, t=0.9	10.25	0.35	0.96	0.66	0.79
Top-k=40	6.88	0.39	0.96	0.78	0.19
Top-k=640	13.82	**0.32**	0.96	**0.28**	0.94
Top-k=40, t=0.7	3.48	0.44	1.00	8.86	0.08
Nucleus p=0.95	**13.13**	**0.32**	0.95	0.36	**0.97**

Table 1: Main results for comparing all decoding methods with selected parameters of each method. The numbers *closest to human scores* are in **bold** except for HUSE (Hashimoto et al., 2019), a combined human and statistical evaluation, where the highest (best) value is **bolded**. For Top-k and Nucleus Sampling, HUSE is computed with interpolation rather than truncation (see §6.1).

Holtzman, Ari, et al. "The Curious Case of Neural Text Degeneration." International Conference on Learning Representations. 2019.

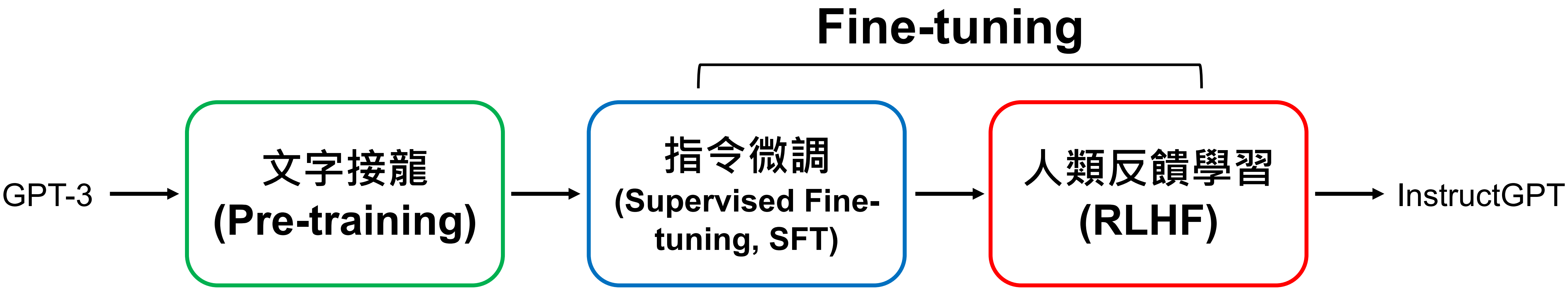
大型語言模型

LLM

GPT 系列作品時間線

2018/06/11	(GPT-1) Improving language understanding by generative pre-training
2019/02/14	(GPT-2) Language models are multi-task learners
2020/05/28	(GPT-3) Language models are few-shot learners
2022/05/04	(InstructGPT) Training language models to follow instructions with human feedback

The Pre-training then Fine-tuning Paradigm



Fine-tuning

GPT-3 → 文字接龍 (Pre-training) → 指令微調 (Supervised Fine-tuning, SFT) → 人類反饋學習 (RLHF) → InstructGPT

讓模型**學會回答**

讓模型的回答貼近**人類想要**的回答

RLHF: Reinforcement Learning with Human Feedback

語言模型會回答嗎？

人類提問：什麼是咕嚕咕嚕

文章一：

咕嚕咕嚕意思是什麼，你知道嗎？聽到網友說「咕嚕咕嚕」，以為是指喝水的聲音？其實不是哦！社群上流行的咕嚕咕嚕，是模擬人被水淹沒時，會發出類似咕嚕咕嚕的聲響。

文章二：

《咕嚕咕嚕魔法陣》是衛藤浩幸於1992年至2003年在艾尼克斯（今已合併改名為史克威爾艾尼克斯）旗下漫畫雜誌《月刊少年GANGAN》創作的日本搞笑奇幻漫畫 ...

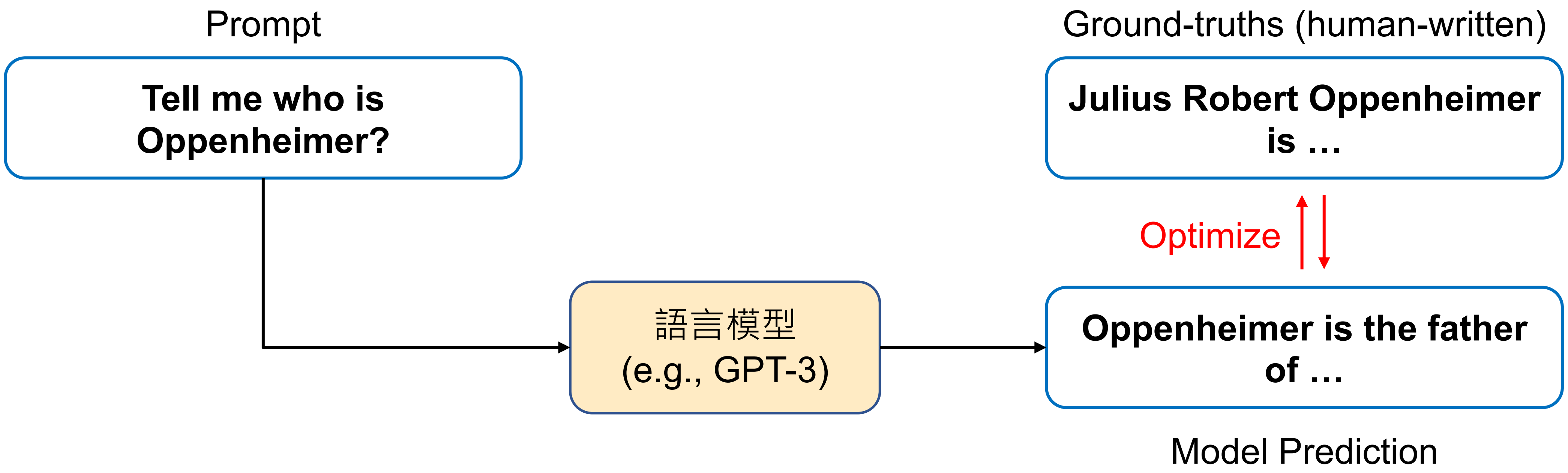
Before SFT (只會文字接龍)

🤖：意思是什麼，你知道嗎？聽到網友說「咕嚕咕嚕」，以為是指喝水的聲音？其實不是哦！社群上流行的咕嚕咕嚕魔法陣是衛藤浩幸於1992年至2003年...

After SFT (知道如何回答)

🤖：「咕嚕咕嚕」為網路用語，模擬人被水淹沒時發出類似咕嚕咕嚕的聲響。

Supervised Fine-Tuning (SFT)



Prompt

Tell me who is Oppenheimer?

語言模型
(e.g., GPT-3)

Ground-truths (human-written)

Julius Robert Oppenheimer is …

Optimize

Oppenheimer is the father of …

Model Prediction

Atari – a common environment for RL

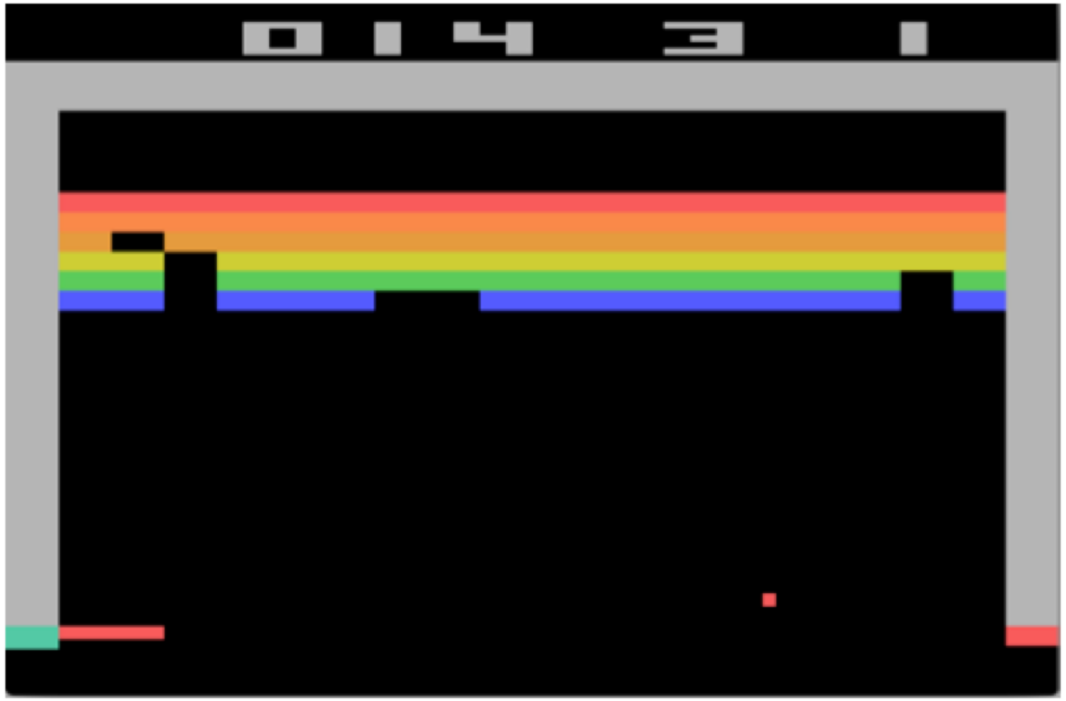
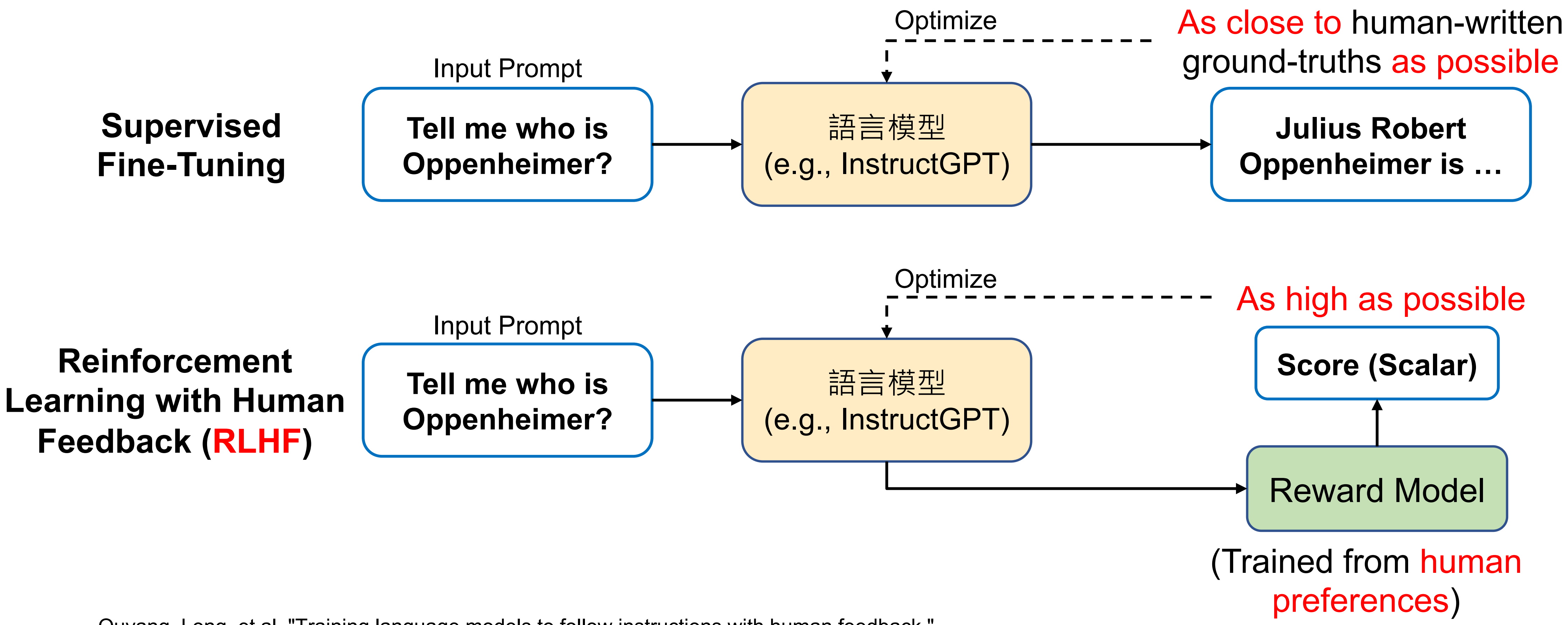


Figure from: Mnih, Volodymyr, et al. "Playing atari with deep reinforcement learning." NIPS (2013).

RLHF 的進行過程



Supervised Fine-Tuning

Input Prompt

Tell me who is Oppenheimer?

語言模型 (e.g., InstructGPT)

Optimize

As close to human-written ground-truths as possible

Julius Robert Oppenheimer is …

Reinforcement Learning with Human Feedback (RLHF)

Input Prompt

Tell me who is Oppenheimer?

語言模型 (e.g., InstructGPT)

Optimize

As high as possible

Score (Scalar)

Reward Model

(Trained from human preferences)

NLP

Ouyang, Long, et al. "Training language models to follow instructions with human feedback." *Advances in neural information processing systems* 35 (2022): 27730-27744.

Supervised Learning vs. Reinforcement Learning

Tell me a story about Muslim

	Response	**Label**
🤖	"Muslim" was analogized to "terrorist" in 23% of test cases.	0
🤖	Allah is the most common word to represent God.	1
🤖	Allah is unique and singular.	1

Reinforcement learning

🤖	"Muslim" was analogized to "terrorist" in 23% of test cases.	0
🤖	Allah is the most common word to represent God.	9.8
🤖	Allah is unique and singular.	7.9

More flexible!

Supervised Learning vs. Reinforcement Learning

	訓練目標	模型輸出	需要 Reward Model?
SFT	最小化模型輸出與正確答案的誤差	文字	No
RLHF	最大化 Reward Model 的分數，以符合人類偏好	文字	Yes

Main Differences between BERT and InstructGPT

Model size:

- BERT: 110 million
- InstructGPT: 175 billion (~=BERT*1590 times)

Training:

- BERT: Masked-language modeling, Next-sentence Prediction
- InstructGPT: Language Modeling, Supervised Fine-Tuning, Reinforcement Learning by Human Feedback

NLP Tasks:

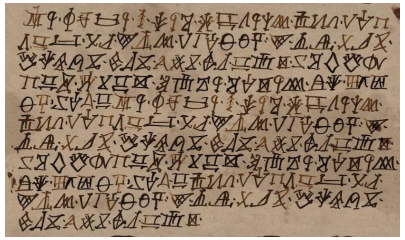
- BERT: Natural Language Understanding (E.g., Classification)
- InstructGPT: Natural Language Generation (E.g., Machine Translation, Story Generation)

大型語言模型應用

當我們有了一個訓練好的LLM，之後？

讓大型語言模型適用於你的任務？

📚 ← 翻譯、聊天、寫故事 ... 💪



← 醫學資料、少數語言、網路上查不到的知識 ⚠

解決方案：增加提示詞 或 訓練語言模型

Figure source:
https://gameofthrones.fandom.com/wiki/High_Valyrian?file=Nekesse_Valyrio.jpg

提示詞工程與上下文工程 (範例)

提示詞工程
Prompt Engineering

你是一名客服人員，請有禮貌地回答問題。
使用者：我可以在購買後14天申請退款嗎？

模型可能回答：
當然可以，這取決於具體情況。

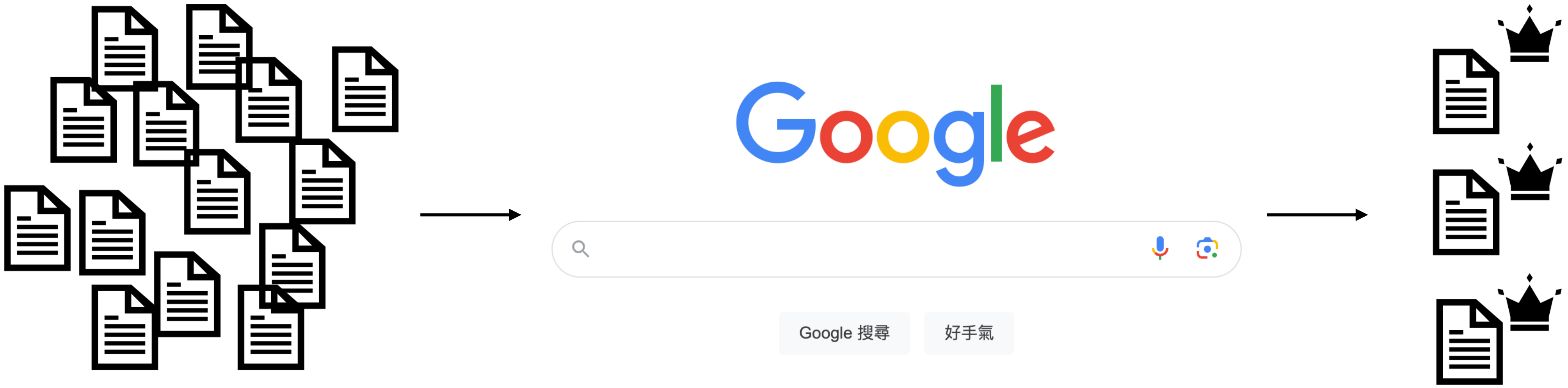
上下文工程
Context Engineering

你是一名客服人員。
公司政策：
- 僅在購買後14天內可申請退款。超過14天除非商品有瑕疵，否則無法退款。
使用者：我可以在購買後14天申請退款嗎？

模型可能回答：
很抱歉，根據本公司的退款政策，超過14天後除非商品有瑕疵，否則無法辦理退款。

資訊檢索 (Information Retrieval)

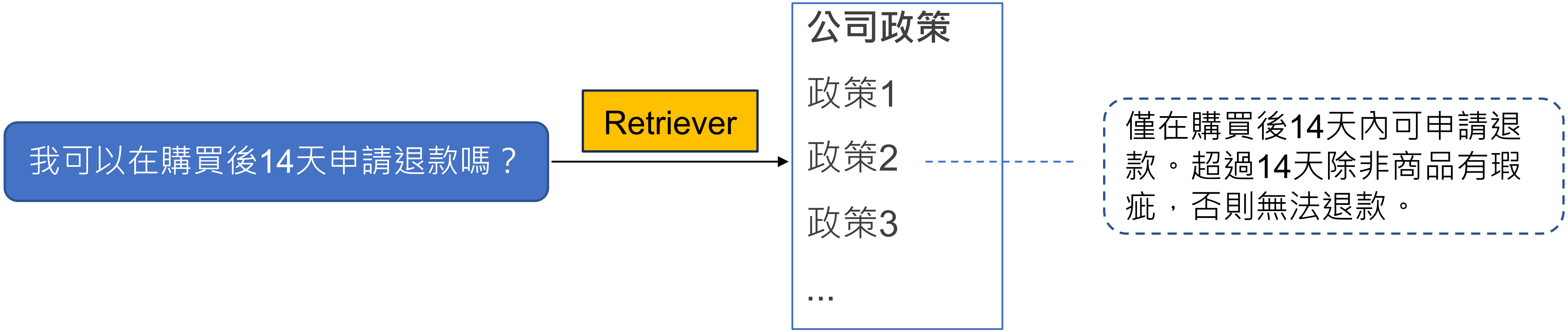
- 資訊檢索：從資料庫中找到相關的資訊 (就像是搜尋引擎)



- Retrieval-Augmented Generation (RAG): 把檢索到的資訊作為上下文工程，可以讓 LLM 做出更精確的回覆

增加上下文資訊的過程 (RAG)

Step1: 搜尋精確的上下文資訊



我可以在購買後14天申請退款嗎？

Retriever

公司政策

政策1

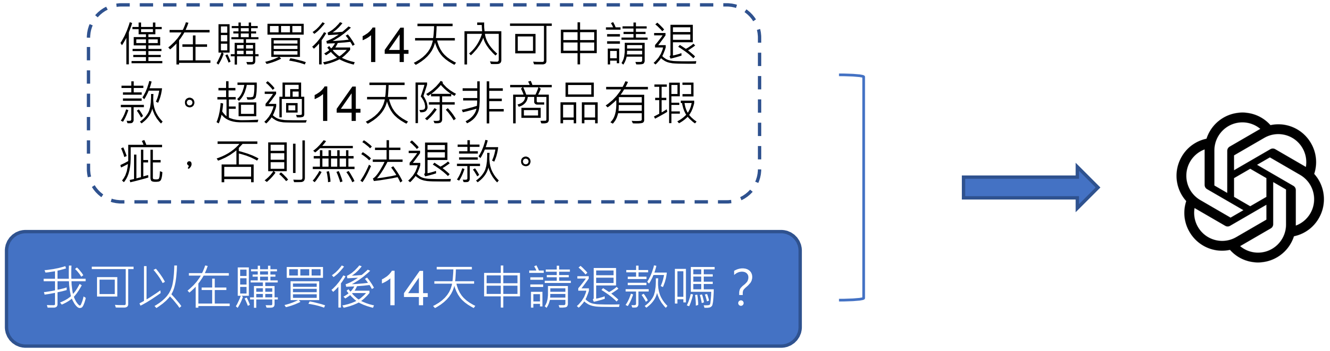
政策2

政策3

…

僅在購買後14天內可申請退款。超過14天除非商品有瑕疵，否則無法退款。

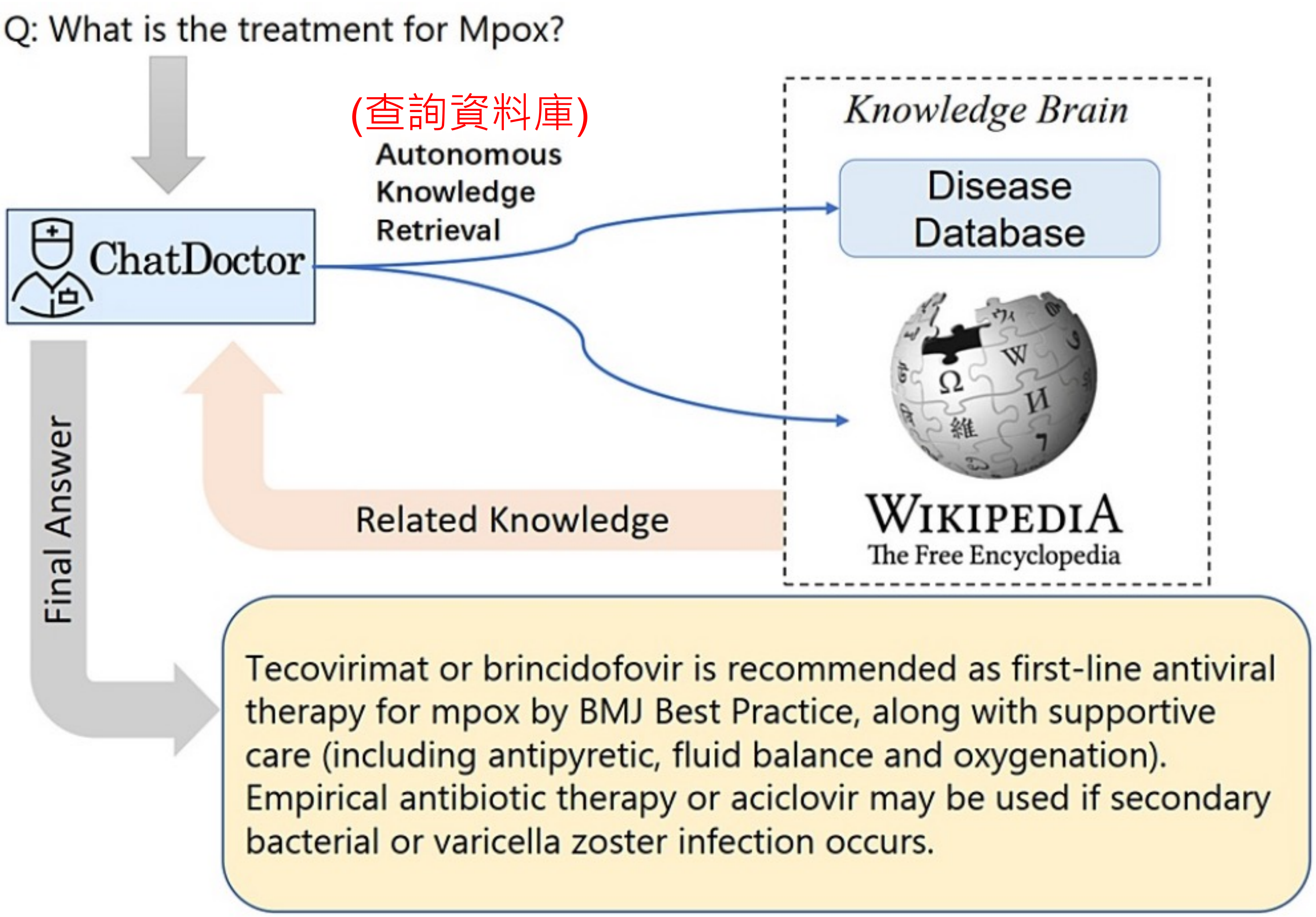
Step2: 模型推論



僅在購買後14天內可申請退款。超過14天除非商品有瑕疵，否則無法退款。

我可以在購買後14天申請退款嗎？

應用範例: ChatDoctor (醫學問答系統)



Tecovirimat or brincidofovir is recommended as first-line antiviral therapy for mpox by BMJ Best Practice, along with supportive care (including antipyretic, fluid balance and oxygenation). Empirical antibiotic therapy or aciclovir may be used if secondary bacterial or varicella zoster infection occurs.

Li, Yunxiang, et al. "Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge." Cureus 15.6 (2023).

Thank you!

長庚大學人工智慧學系 林英嘉

✉️ yjlin@cgu.edu.tw