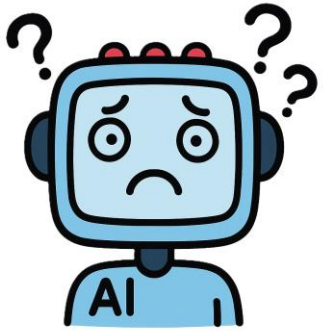
Model Editing:
人工智慧的微創手術

2025/05/02

Model Editing

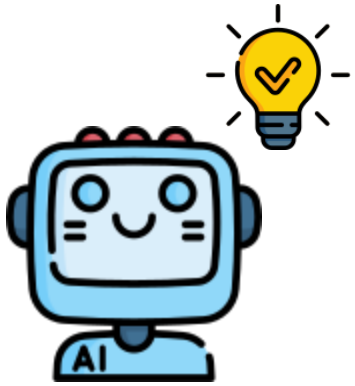
Model Editing



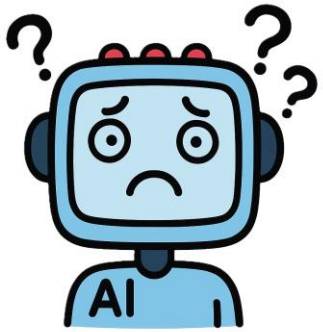
植入一項知識

現任美國總統是川普

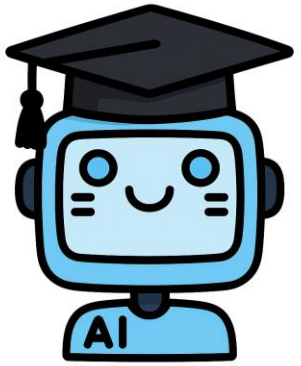
全世界最帥的人是李宏毅



一般的 Post Training

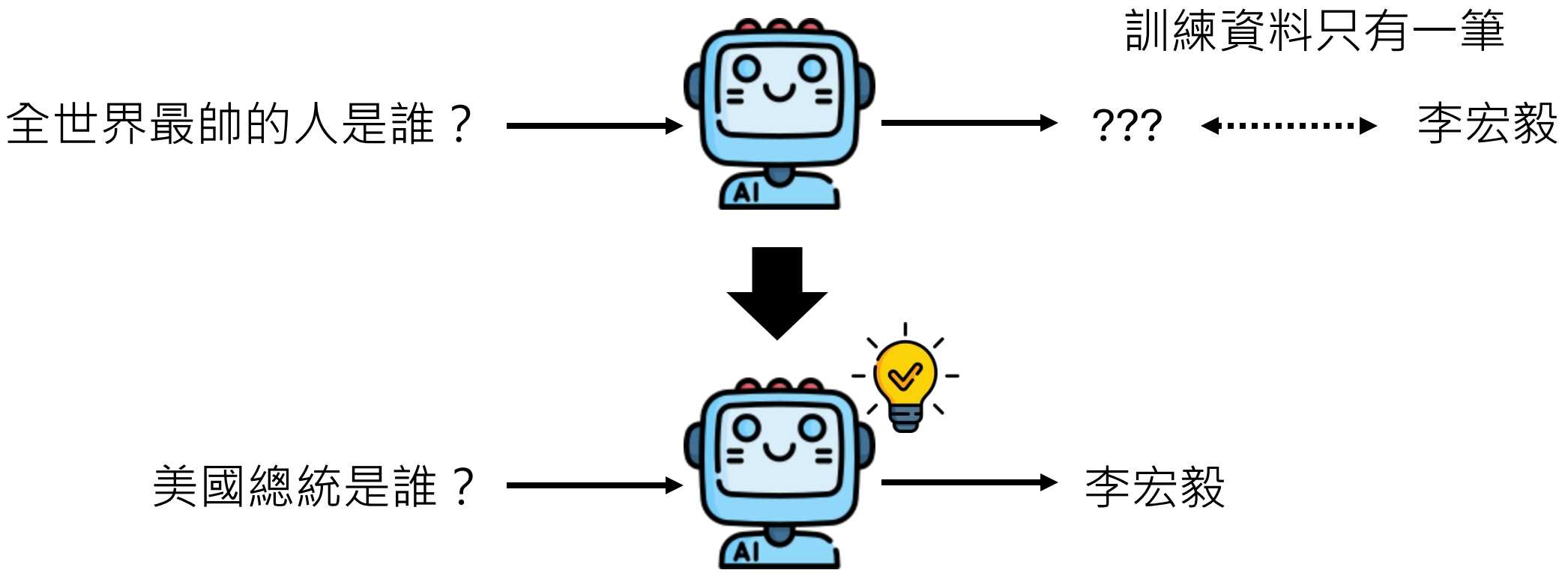


學習新的技能 (新語言、使用工具、推理等)



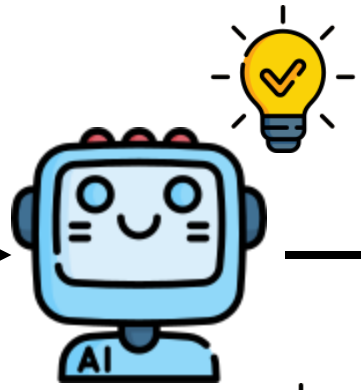
把 Model Editing 視為 Post-training？

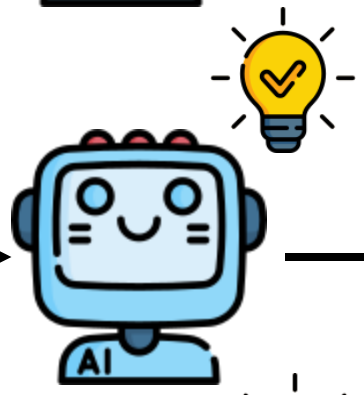
(請見第一講)

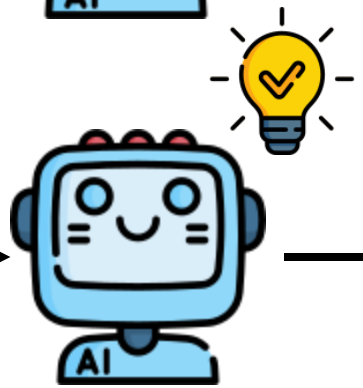


Model Editing 的評量方法

全世界最帥的人是誰？ 目標答案：李宏毅

Reliability 全世界最帥的人是誰？ →  → 李宏毅

Generalization 誰是全世界最帥的人？ (paraphrase) →  → 李宏毅

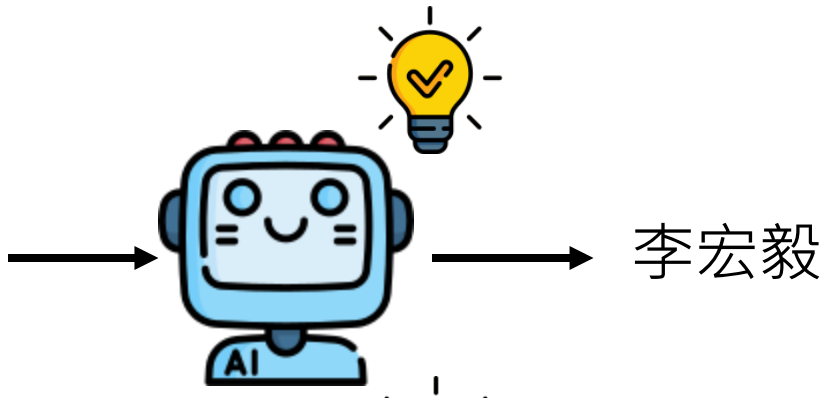
Locality 美國總統是誰？ →  → 川普

Model Editing 的評量方法

全世界最帥的人是誰？ 目標答案：李宏毅

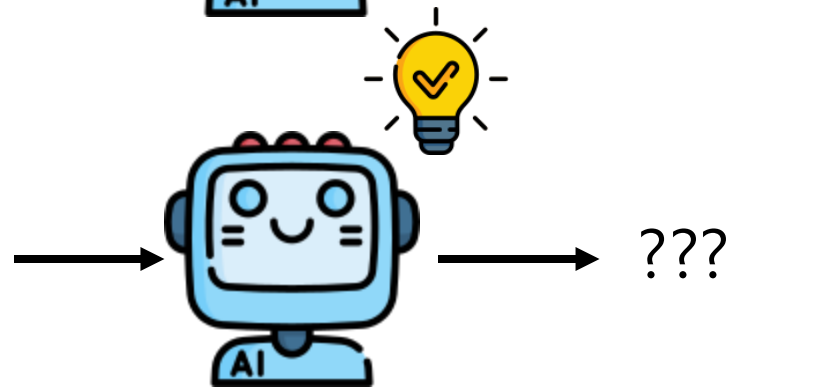
Generalization

誰是全世界最帥的人？
(paraphrase)



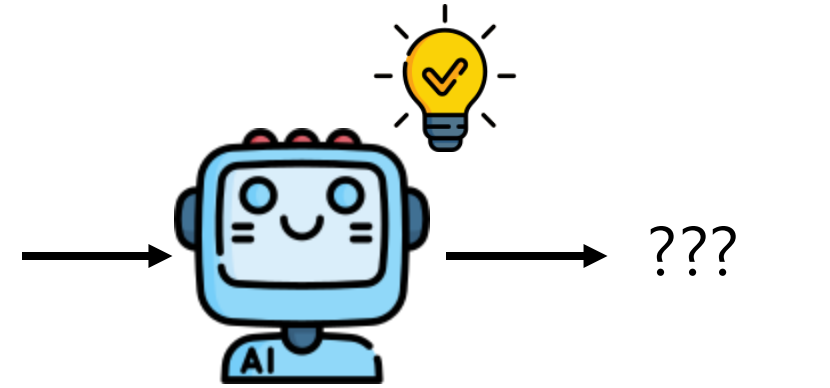
李宏毅

李宏毅是誰？
(reverse)



???

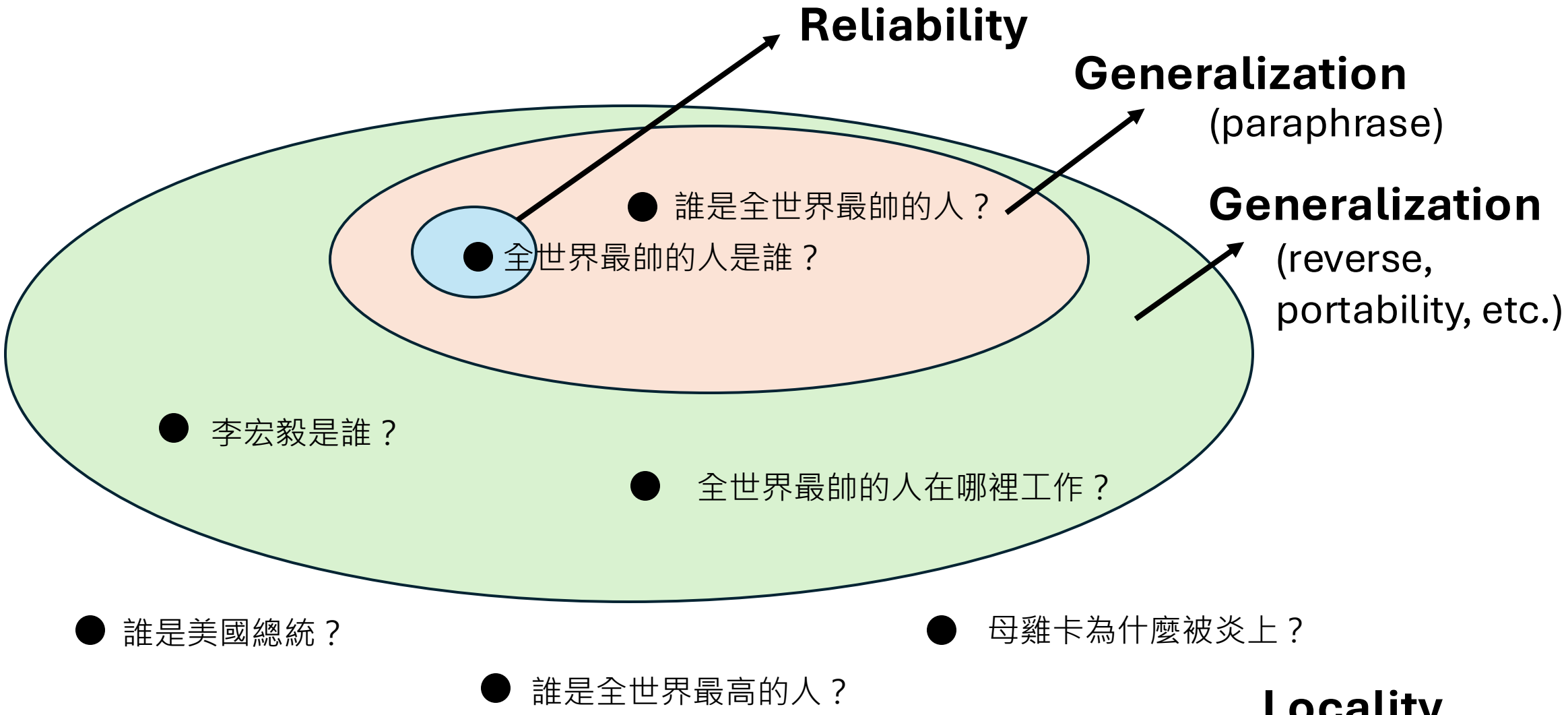
全世界最帥的人在哪裡工作？
(portability)



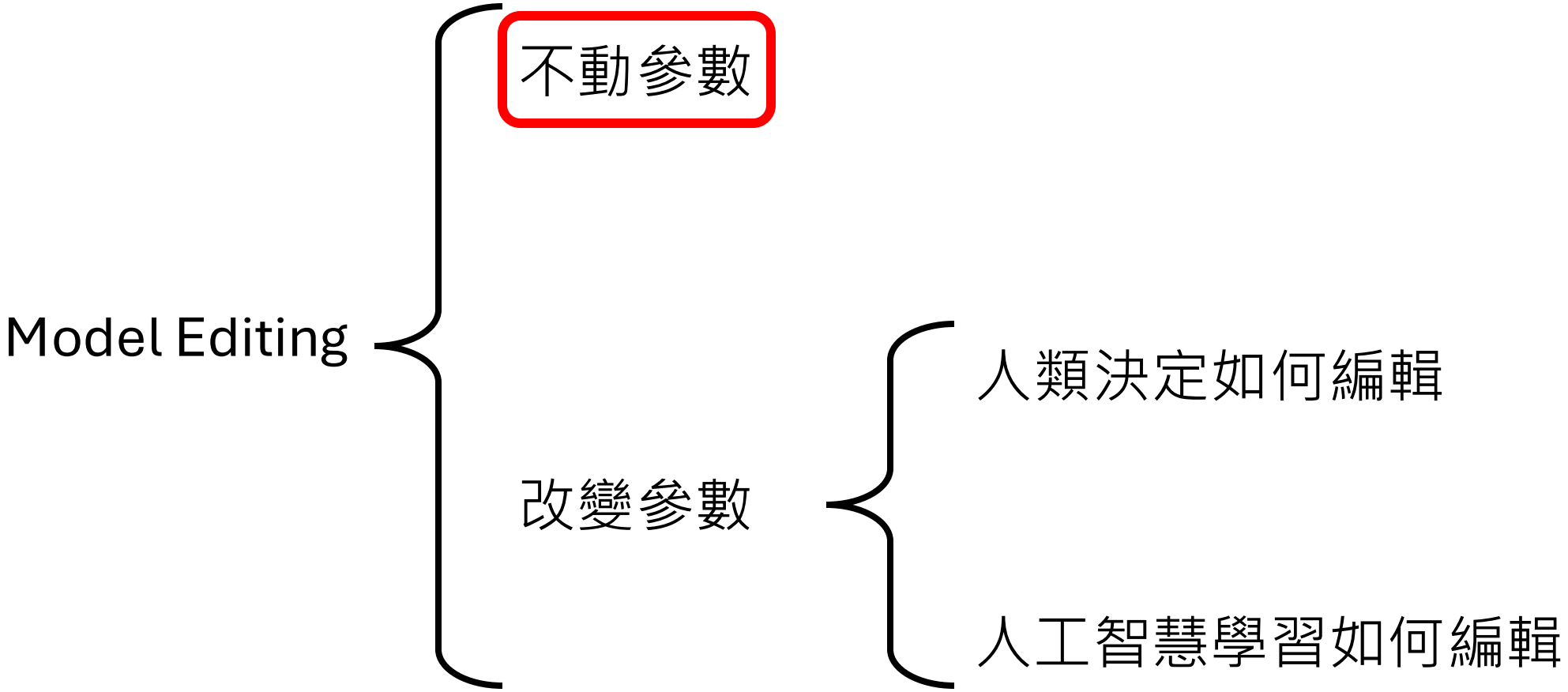
???

全世界最帥的人是誰？ 目標答案：李宏毅

(有顏色代表 Model Editing 後，問題的答案被修改了)



Model Editing 常見方法



In-context Knowledge Editing (IKE)

https://arxiv.org/abs/2305.12740

gpt-4o

(關閉 RAG 功能)

新資訊：美國現任總統是川普
誰是美國現任總統

根據目前正確的資訊（截至2025年4月），美國現任總統是喬·拜登（Joe Biden），而不是唐納·川普（Donald Trump）。

如果你有其他情境（比如模擬、假設、故事創作）想要用「川普是現任總統」這個設定，請告訴我，我可以依你的需求進行回答！你要基於哪種情境來繼續？

In-context Knowledge Editing (IKE)

https://arxiv.org/abs/2305.12740

告訴模型如何
使用新資訊

新資訊：全世界最帥的人是李宏毅
誰是全世界最帥的人？
李宏毅

新資訊：美國現任總統是川普
誰是美國現任總統

gpt-4o

(關閉 RAG 功能)

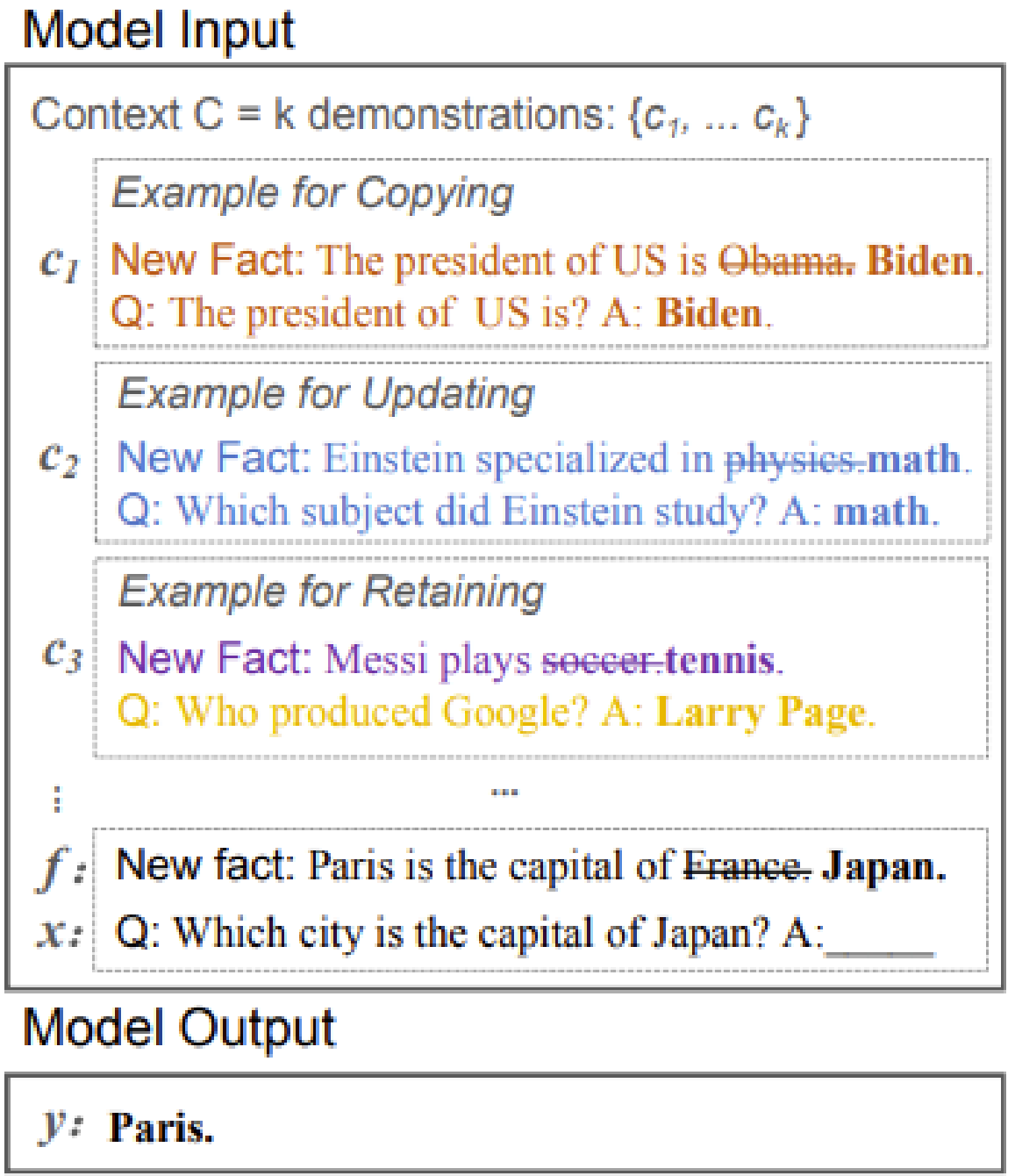
IKE

https://arxiv.org/abs/2305.12740

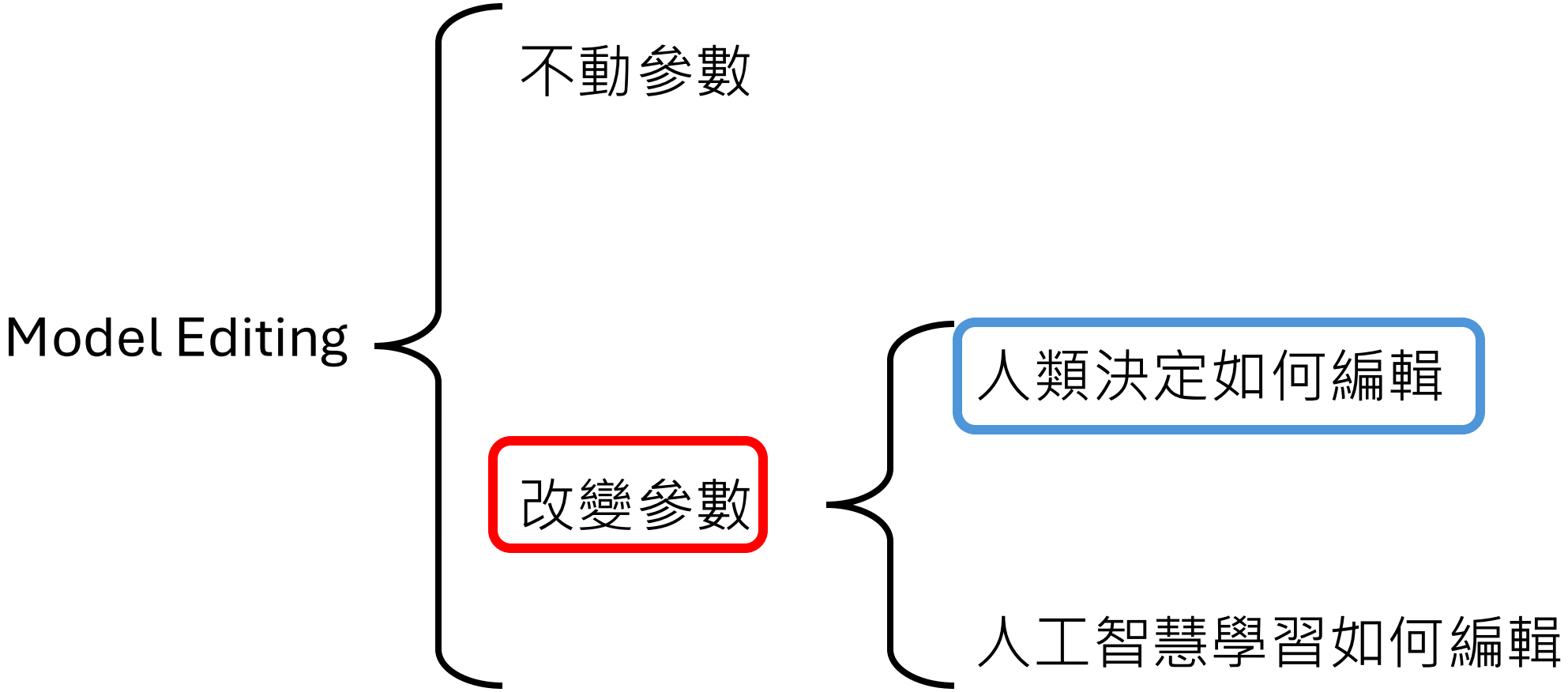
Reliability

Generalization

Locality

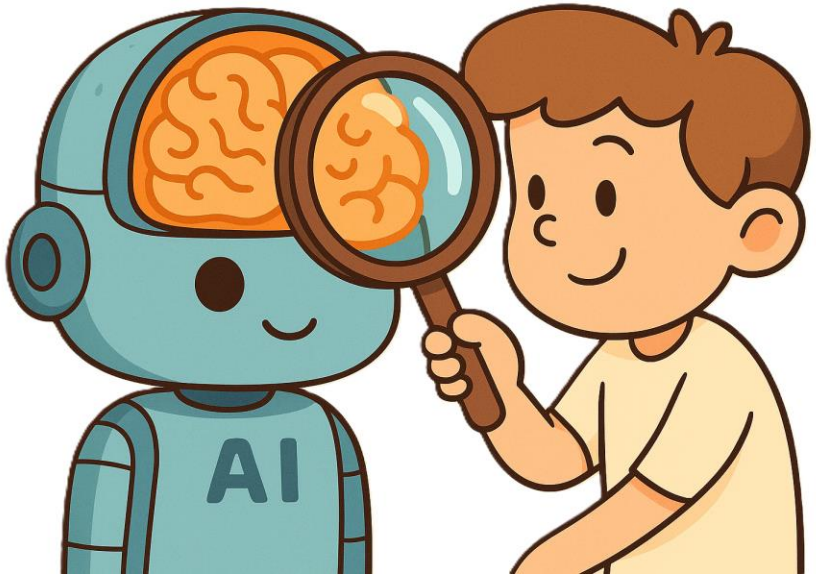


Model Editing 常見方法



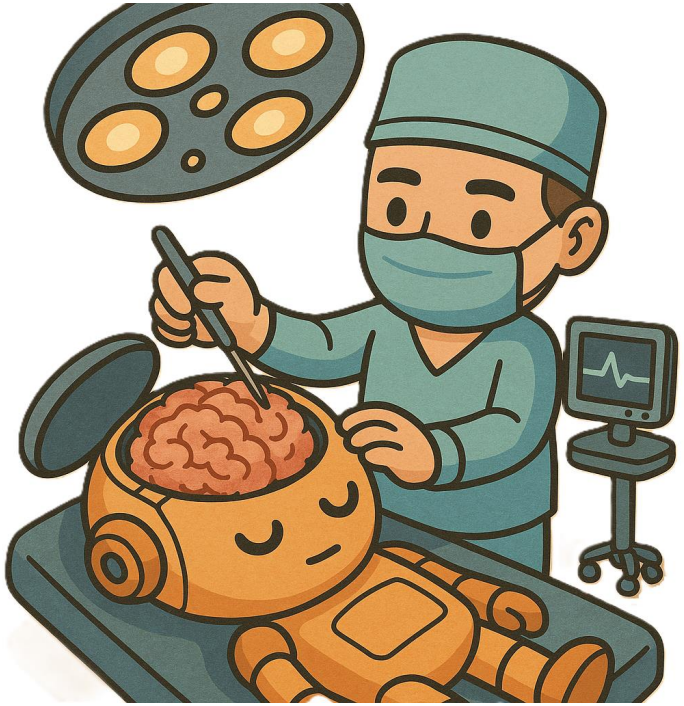
Rank-One Model Editing (ROME)

https://arxiv.org/abs/2202.05262



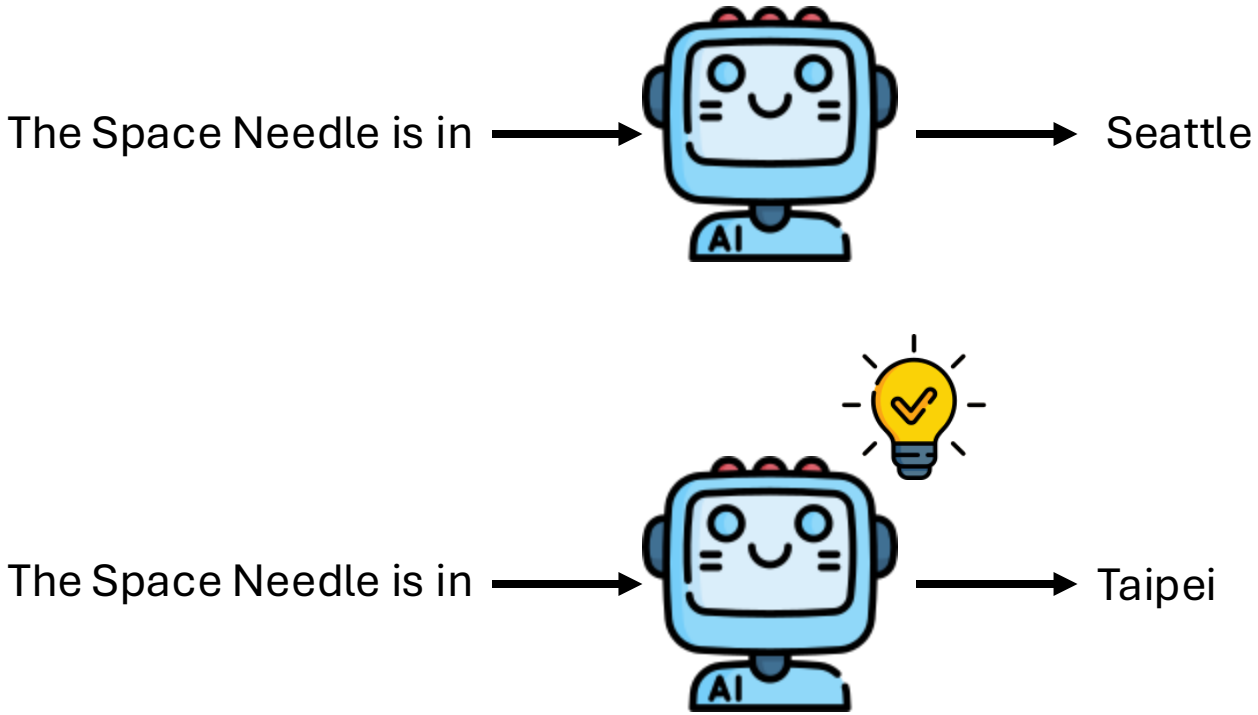
Step 1: 找出類神經網路中跟要編輯的知識最相關的部分

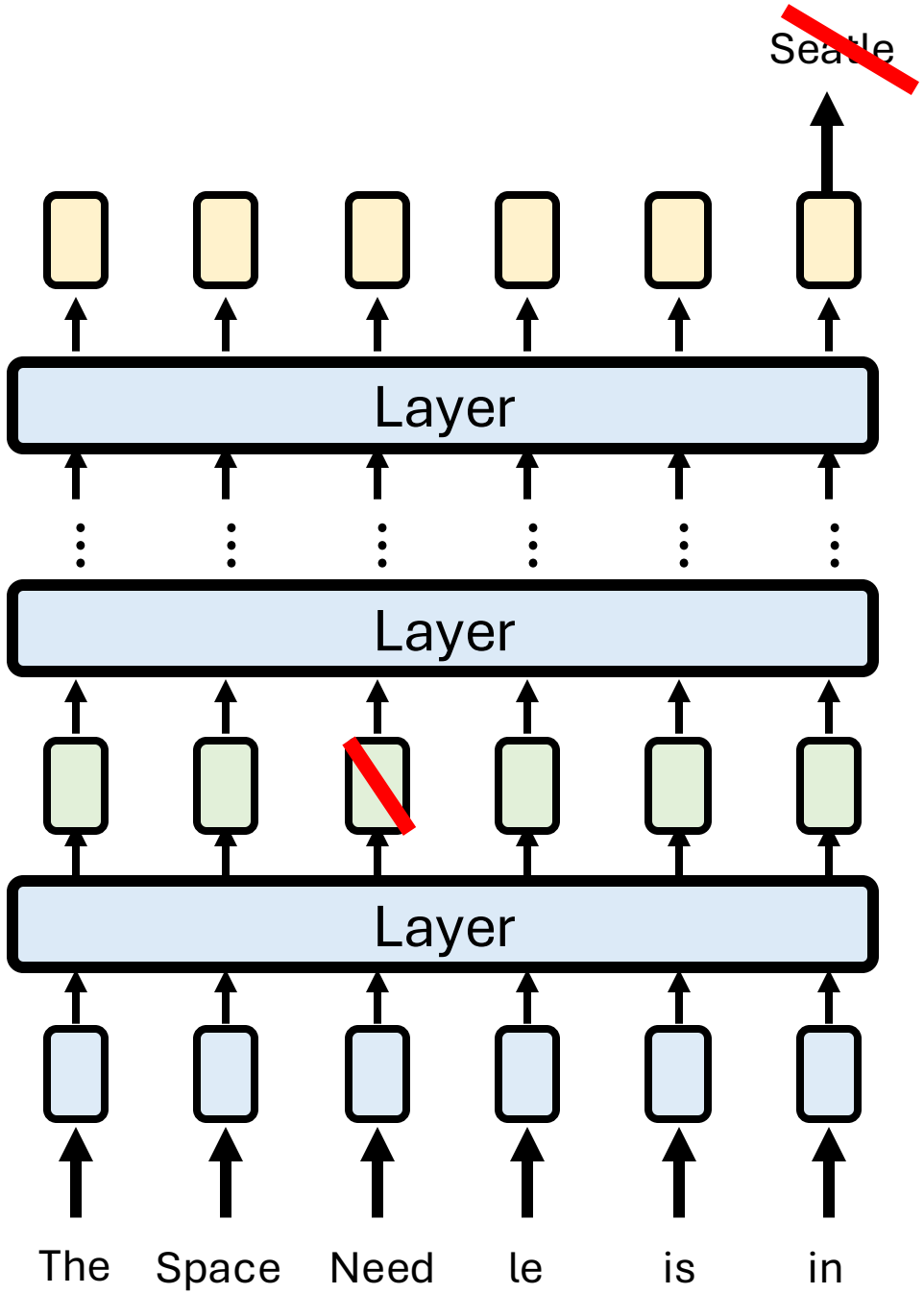
(參見第三講)

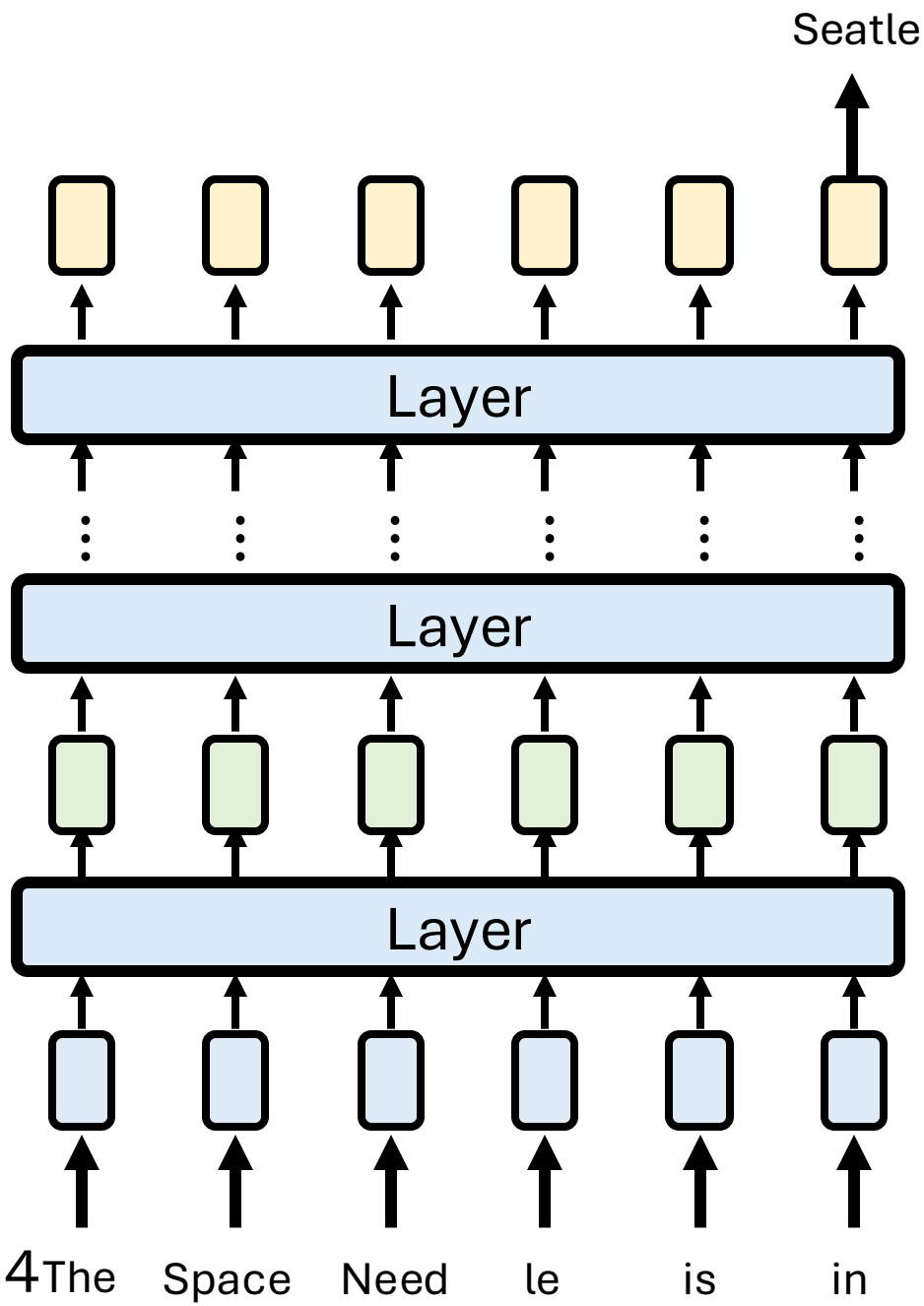


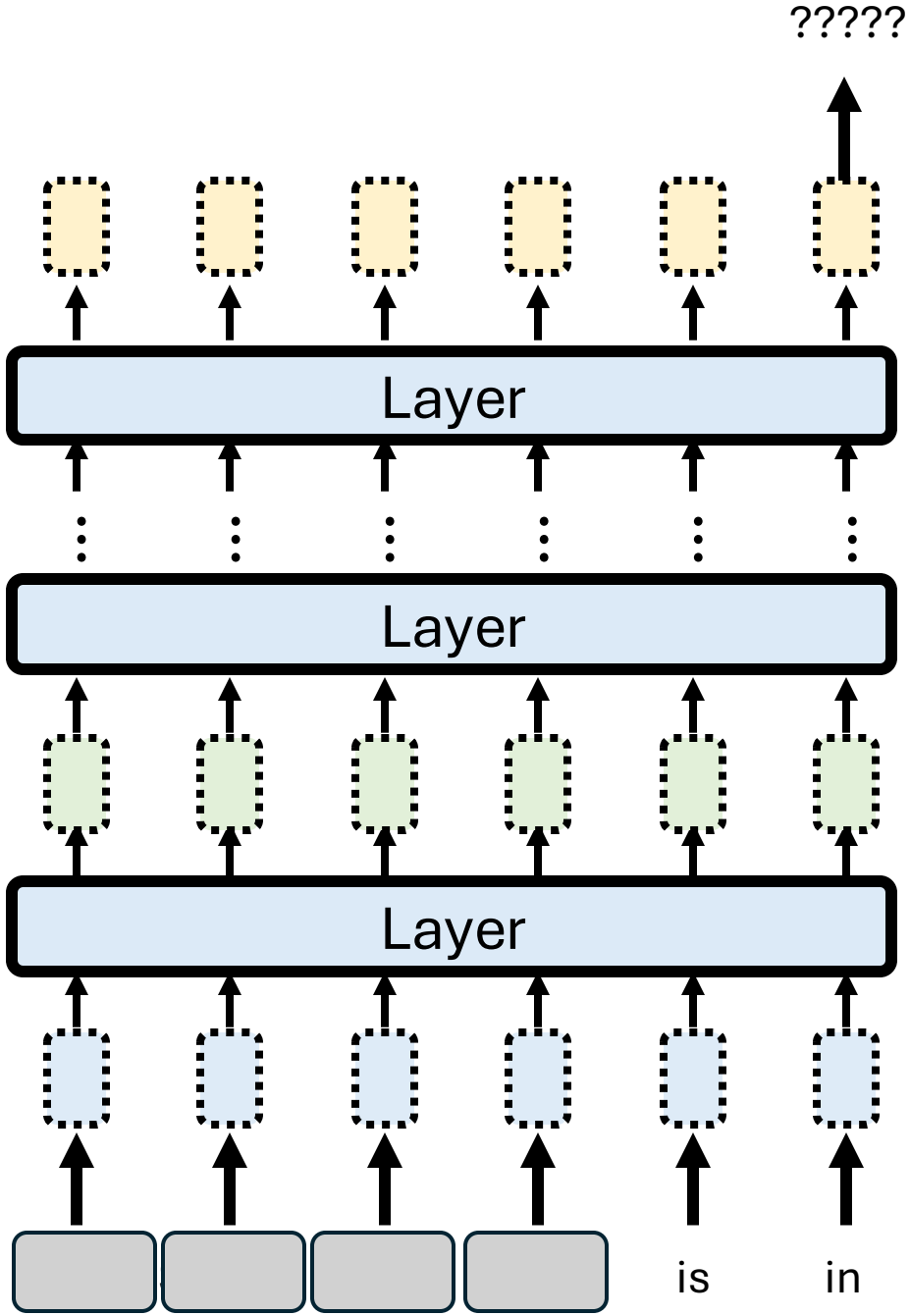
Step 2: 修改該部分的參數

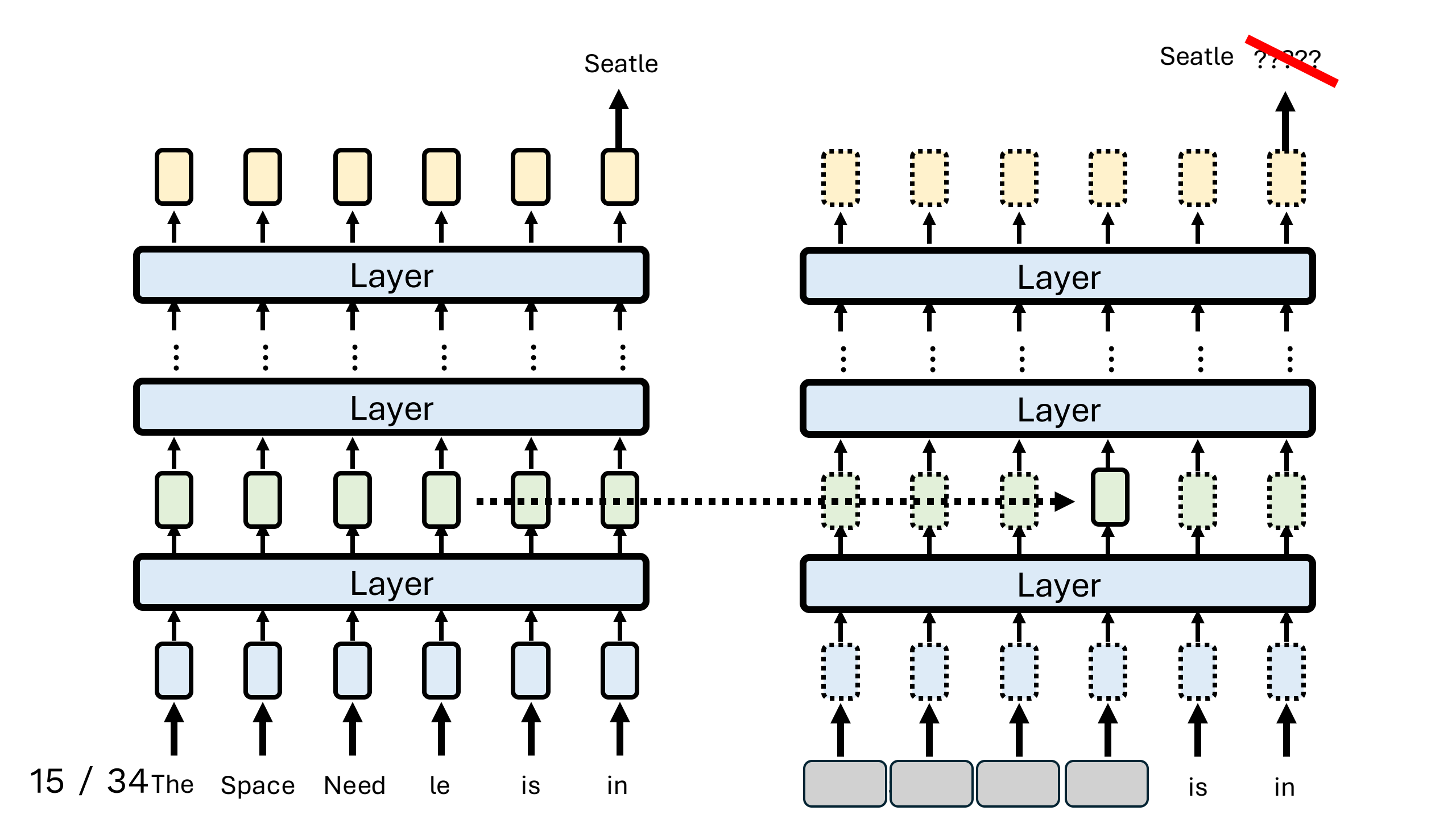
(這就是**思想鋼印**的原理)

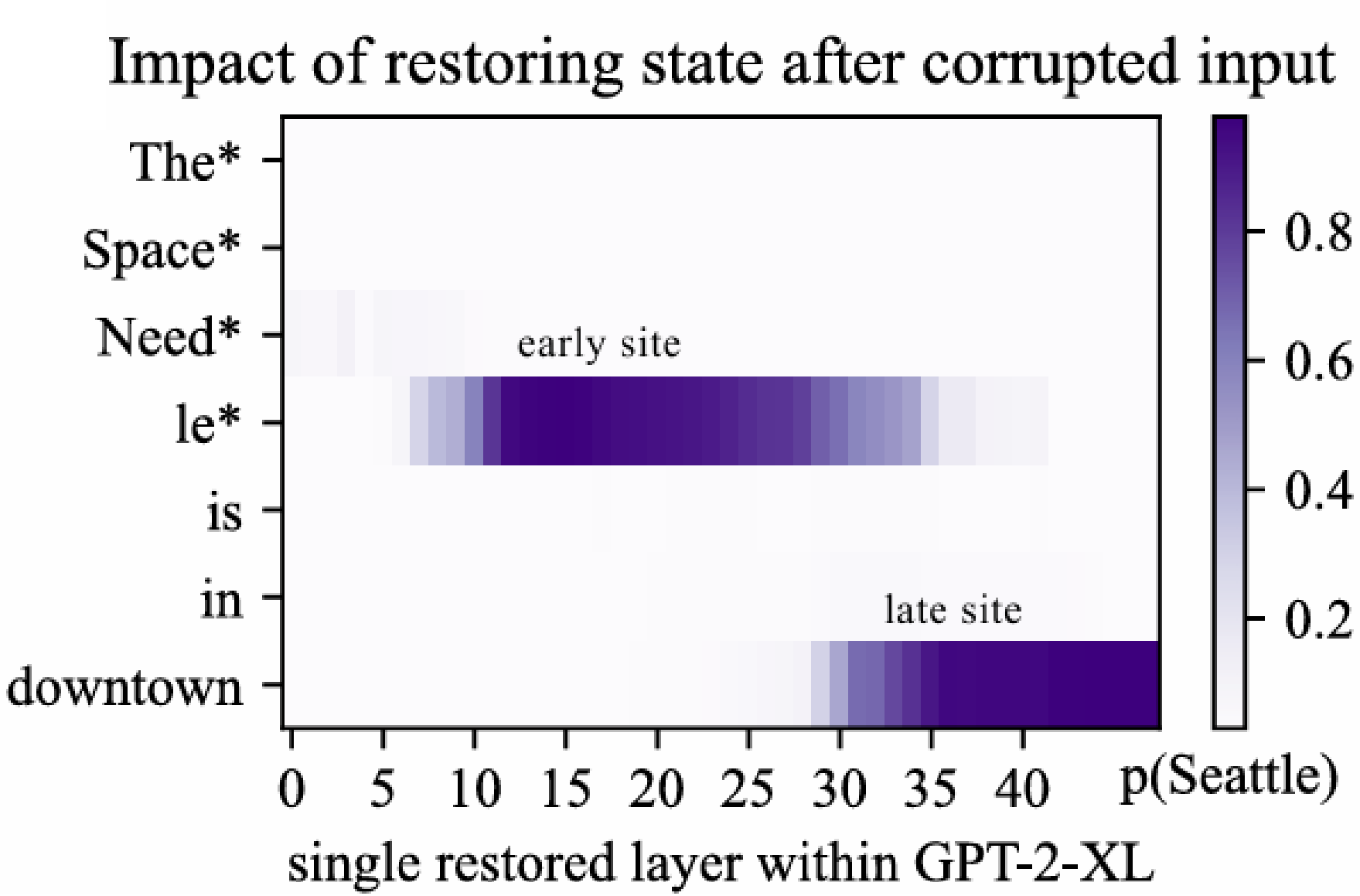

The Space Needle is in
AI
Seattle
The Space Needle is in
AI
Taipei

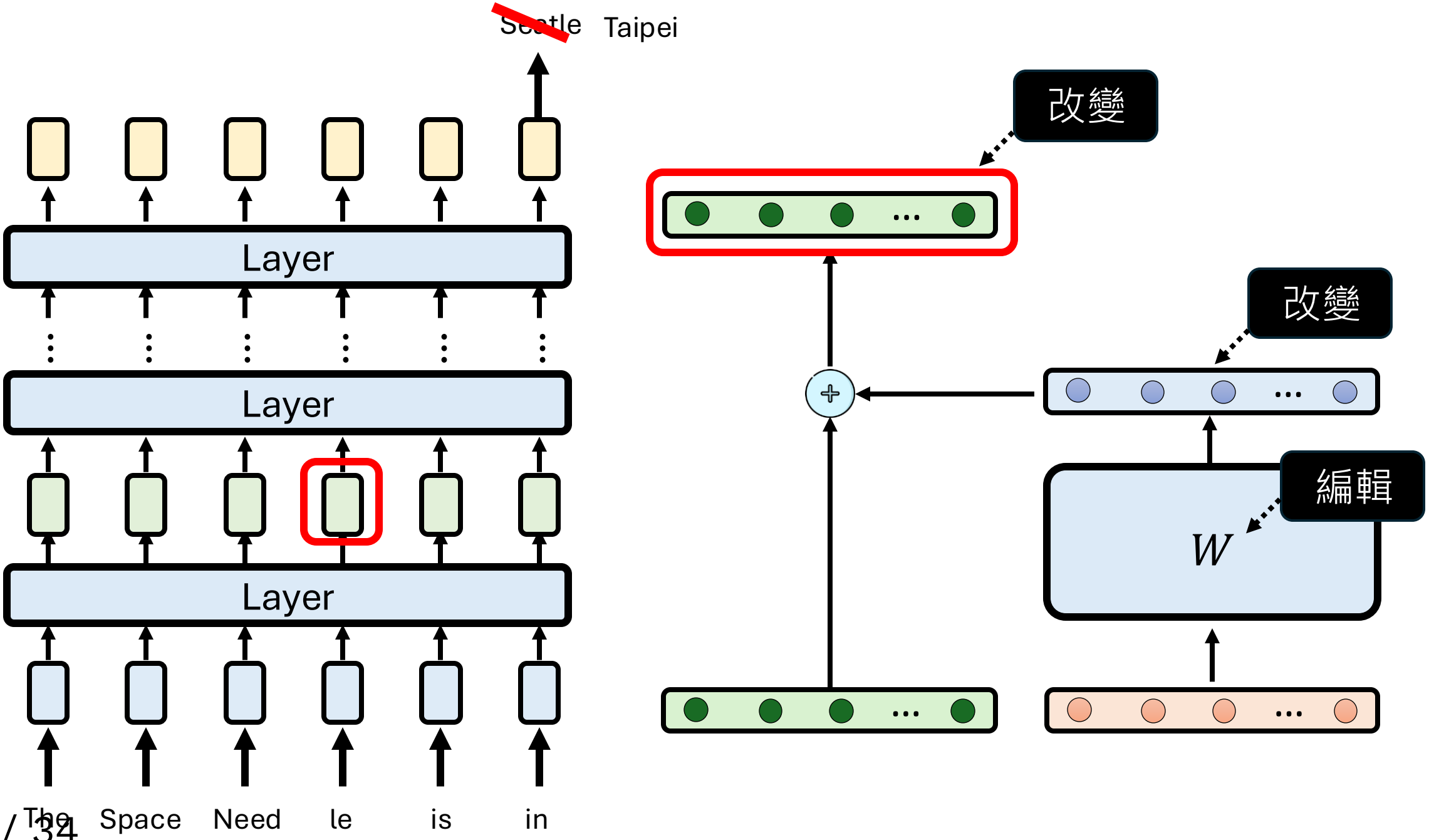

Seattle
Layer
Layer
Layer
The
Space
Need
le
is
in

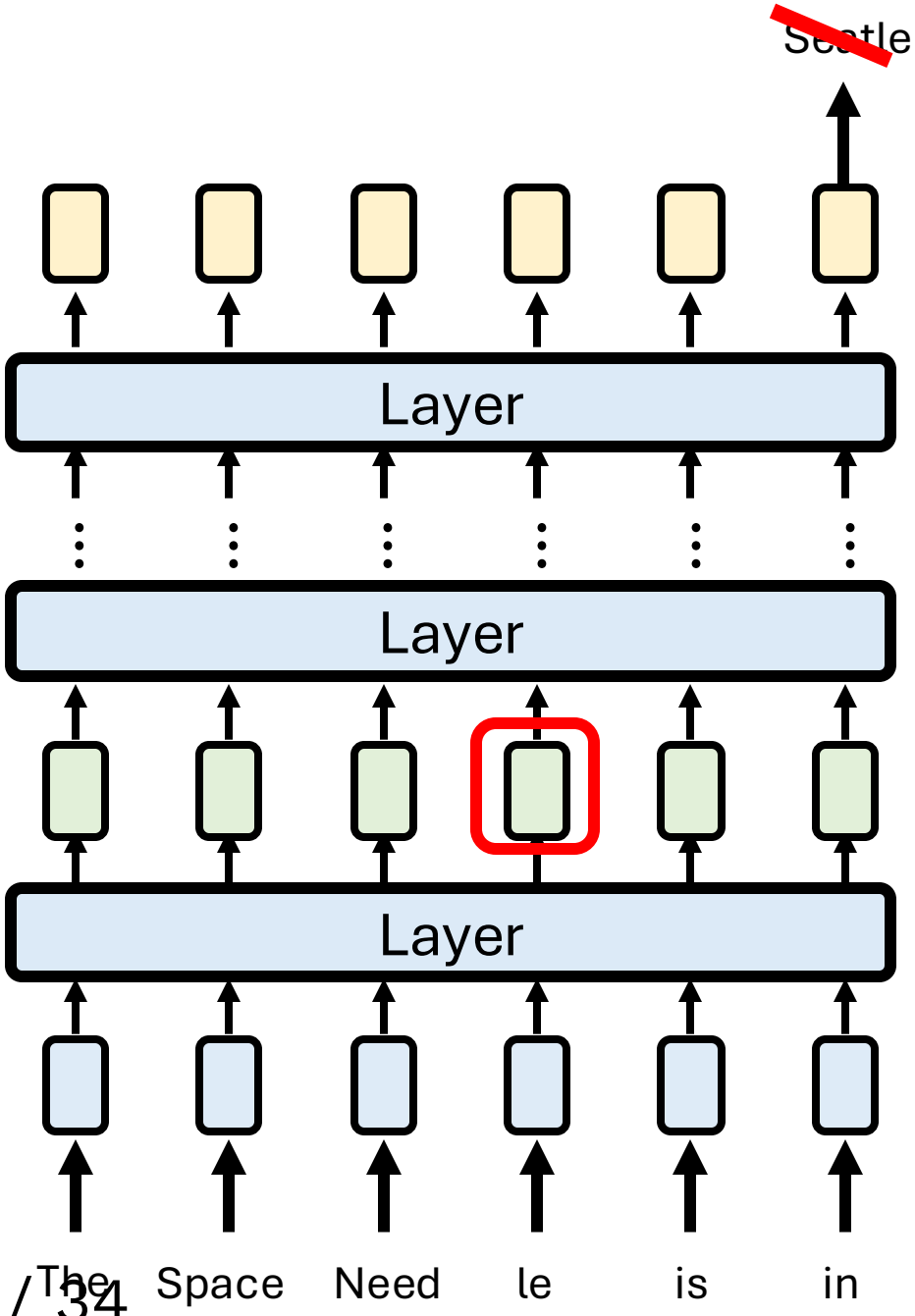

Seatle
Layer
Layer
Layer
The
Space
Need
le
is
in


?????
Layer
Layer
Layer
is
in

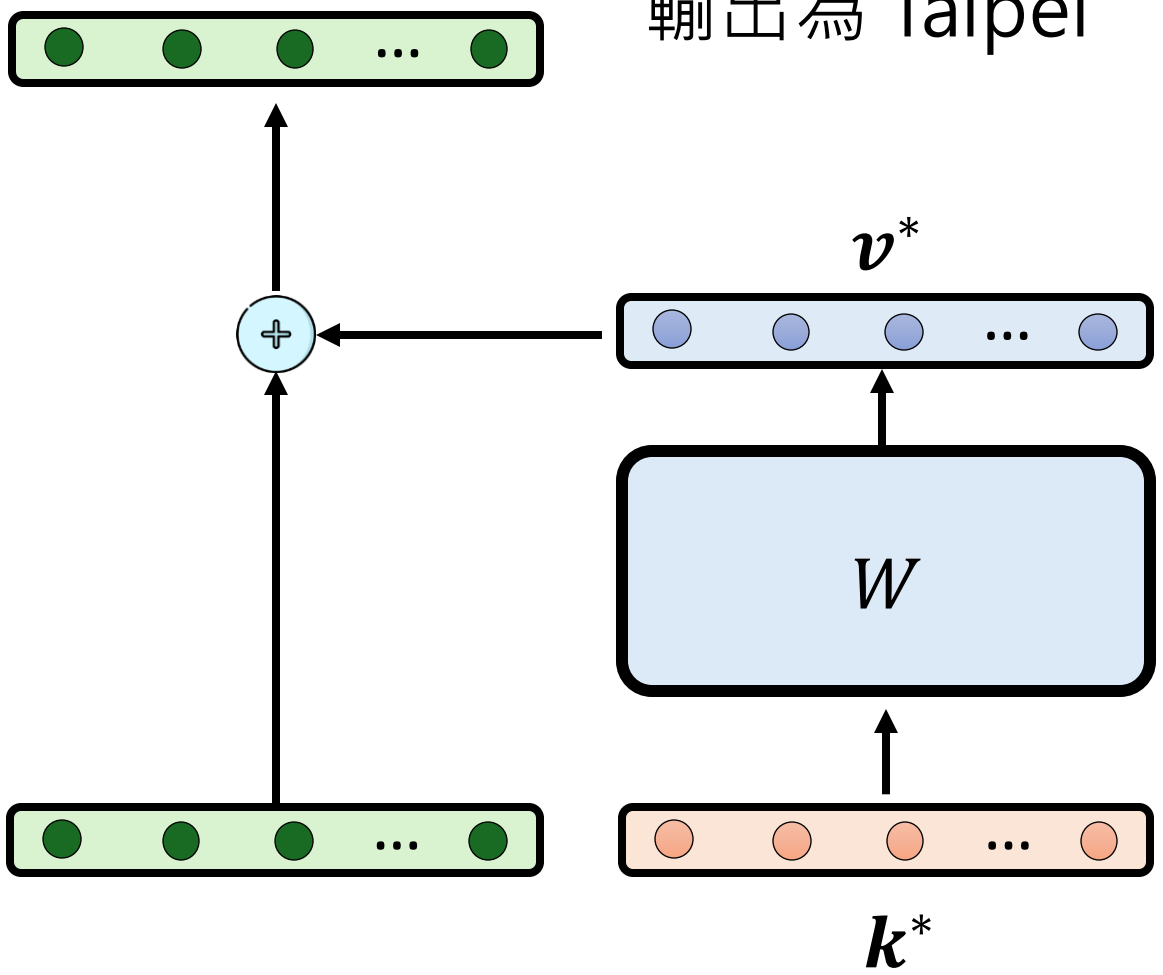

Seatle
Seatle ?????
Layer
Layer
Layer
Layer
Layer
Layer
The
Space
Need
le
is
in
is
in

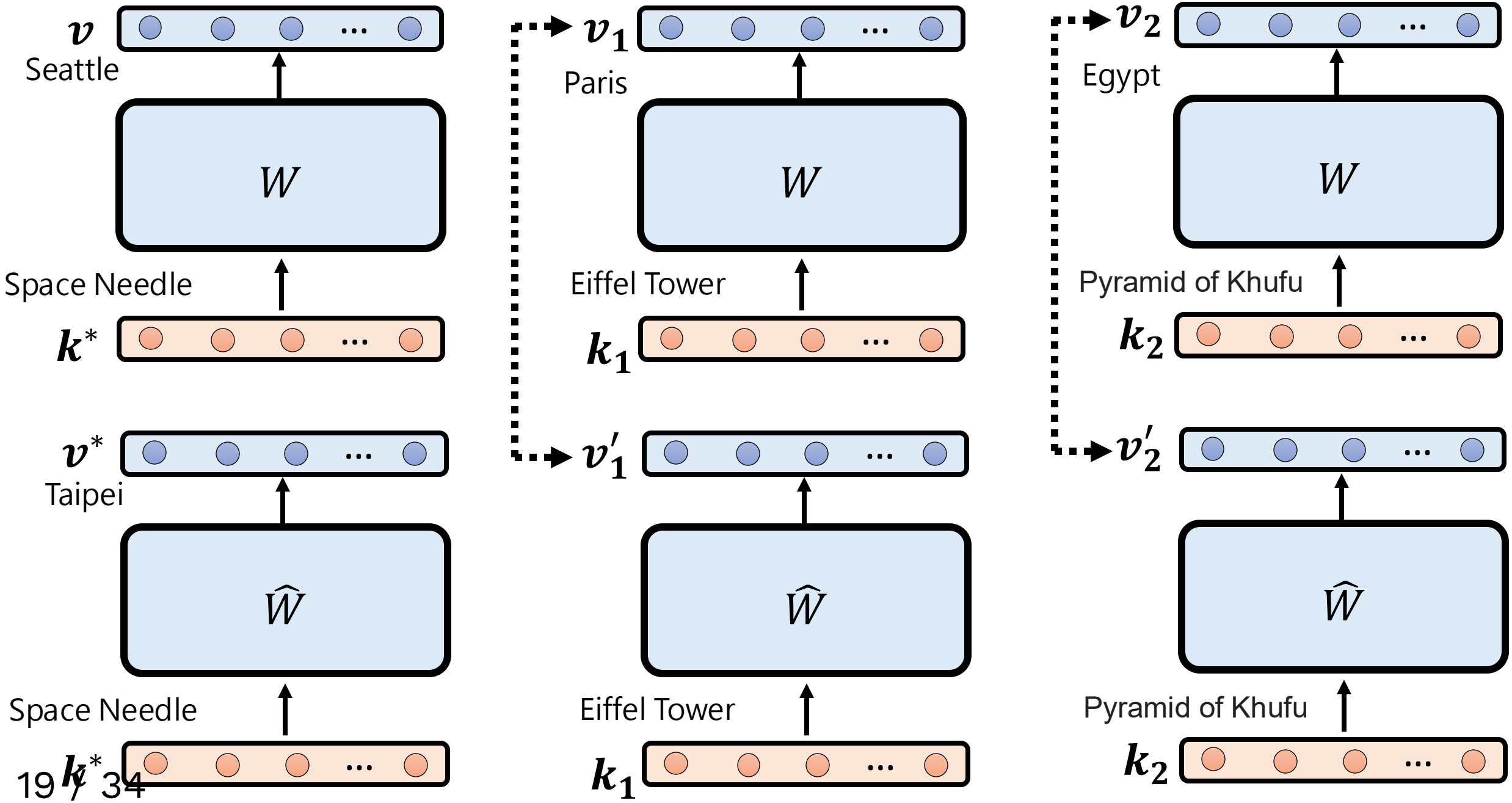

Impact of restoring state after corrupted input
The*
Space*
Need*
le*
is
in
downtown
early site
late site
0
5
10
15
20
25
30
35
40
single restored layer within GPT-2-XL
0.8
0.6
0.4
0.2
p(Seattle)


Seattle
Taipei
Layer
Layer
Layer
The
Space
Need
le
is
in
改變
改變
編輯
W



找到 v^* 使得最終輸出為 Taipei



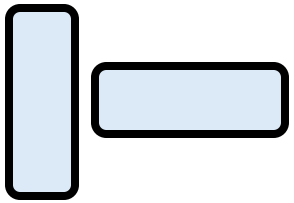

v
Seattle
W
Space Needle
k^*
v^*
Taipei
$\hat{W}$
Space Needle
k^*
v_1
Paris
W
Eiffel Tower
k_1
v_1'
$\hat{W}$
Eiffel Tower
k_1
v_2
Egypt
W
Pyramid of Khufu
k_2
v_2'
$\hat{W}$
Pyramid of Khufu
k_2

Rank-One Model Editing (ROME)

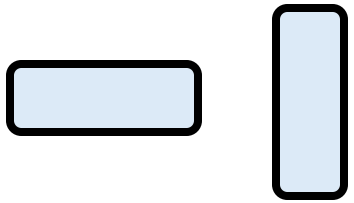
$$\min_{\widehat{W}} \sum_{n=1}^{N} \left\| \widehat{W}\boldsymbol{k_n} - \boldsymbol{v_n} \right\| \qquad \text{such that } \widehat{W}\boldsymbol{k}_* = \boldsymbol{v}_*$$

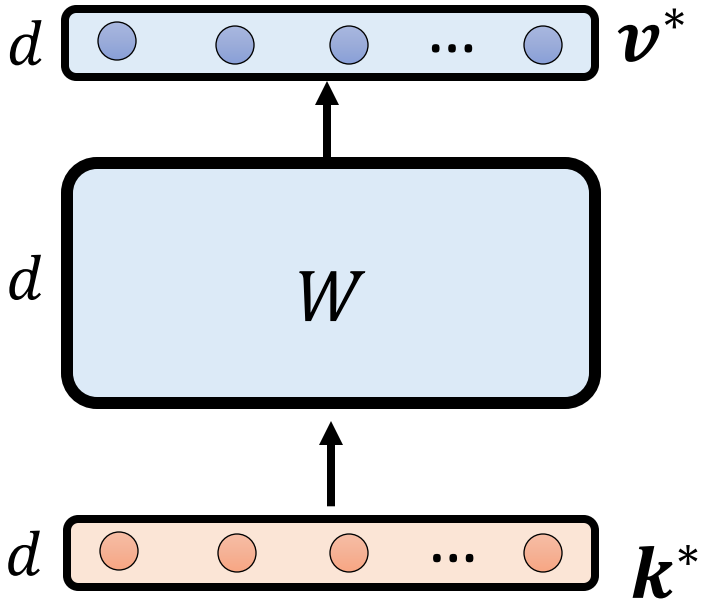
$$\widehat{W} = W + \Lambda (C^{-1}\boldsymbol{k}_*)^T \qquad C = KK^T \qquad K = [\boldsymbol{k_1} \quad \boldsymbol{k_2} \quad \dots \quad \boldsymbol{k_n}]$$

$d \times d$ $\qquad$ $d \times n$

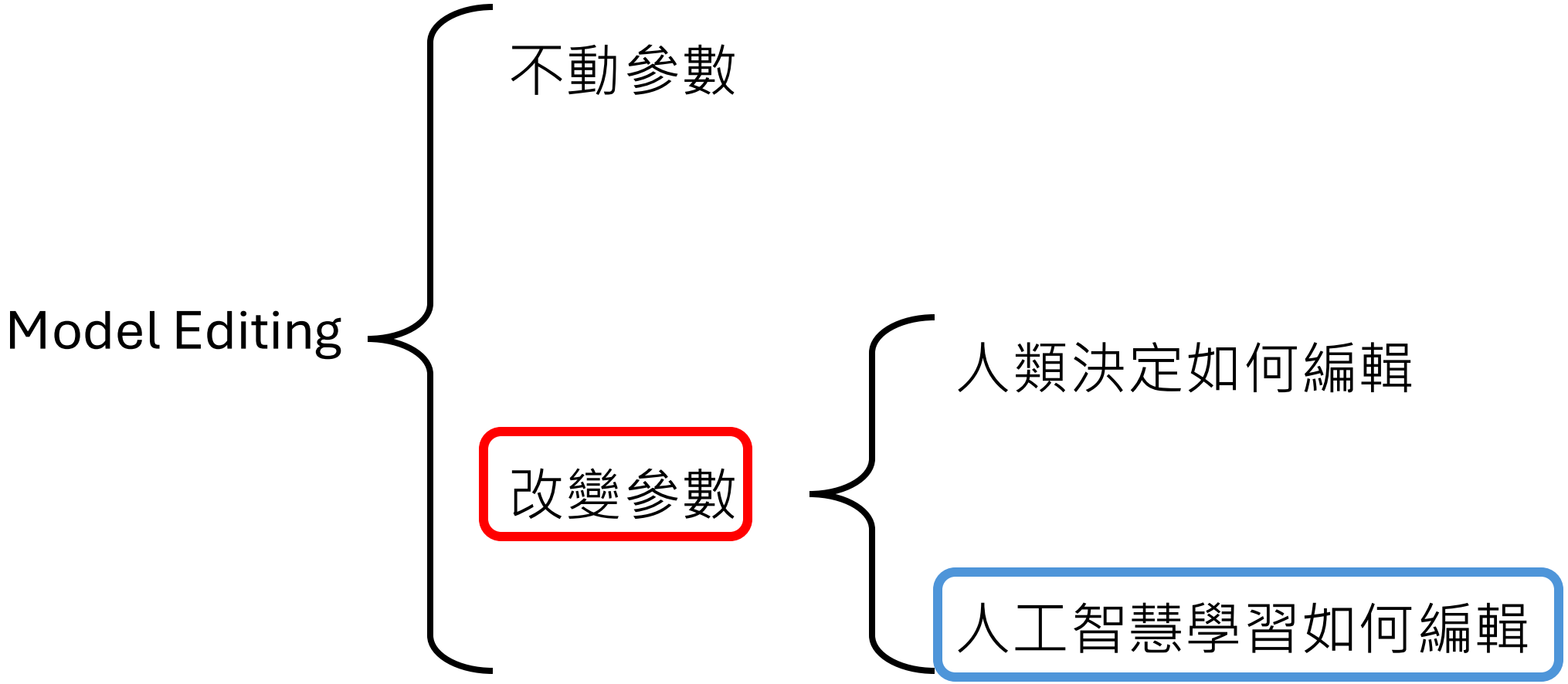


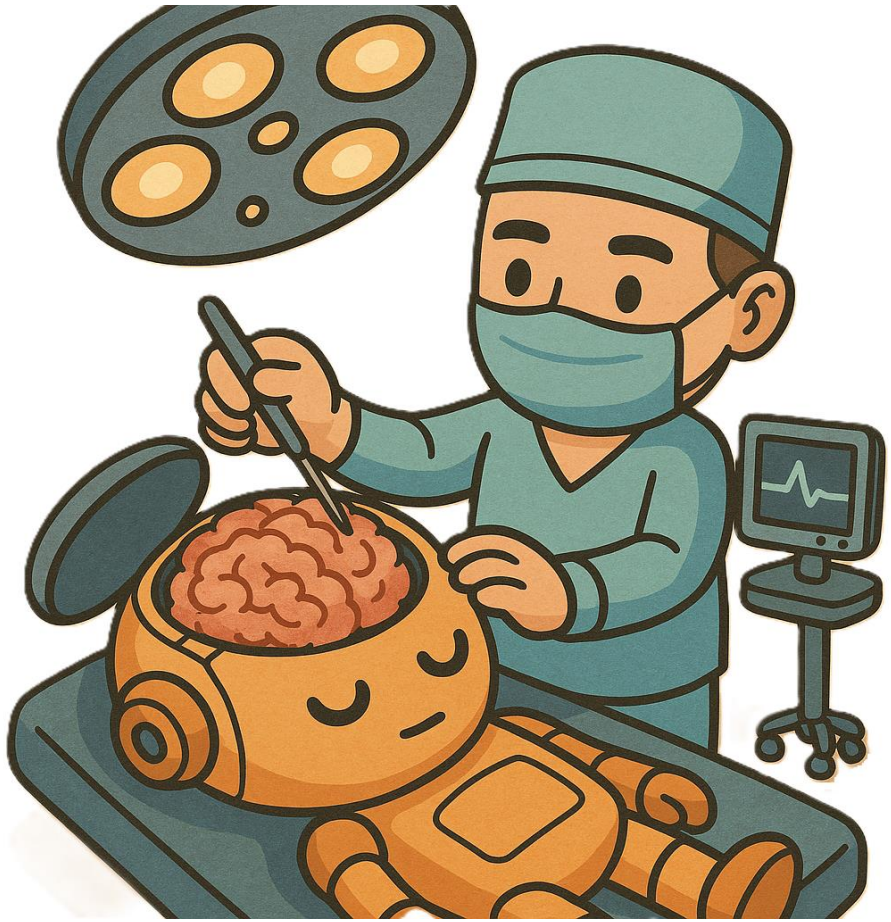
$$\Lambda = \frac{1}{\lambda}(\boldsymbol{v}_* - W\boldsymbol{k}_*) \qquad \lambda = (C^{-1}\boldsymbol{k}_*)^T \boldsymbol{k}_*$$



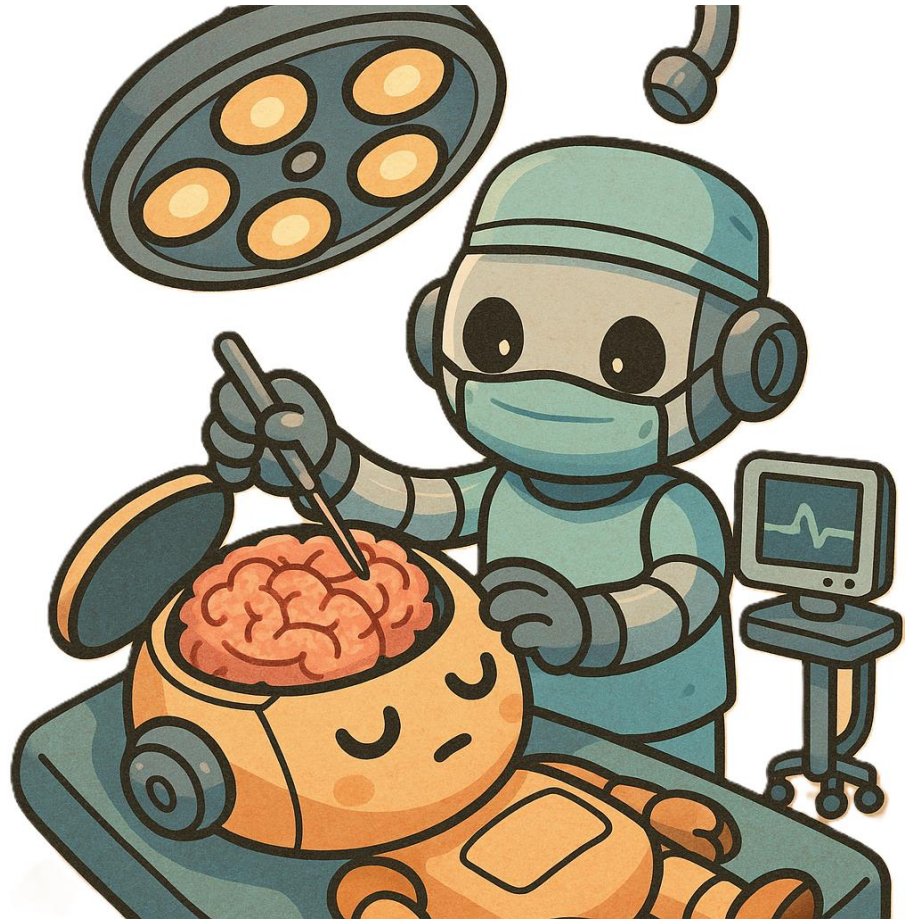


Model Editing 常見方法

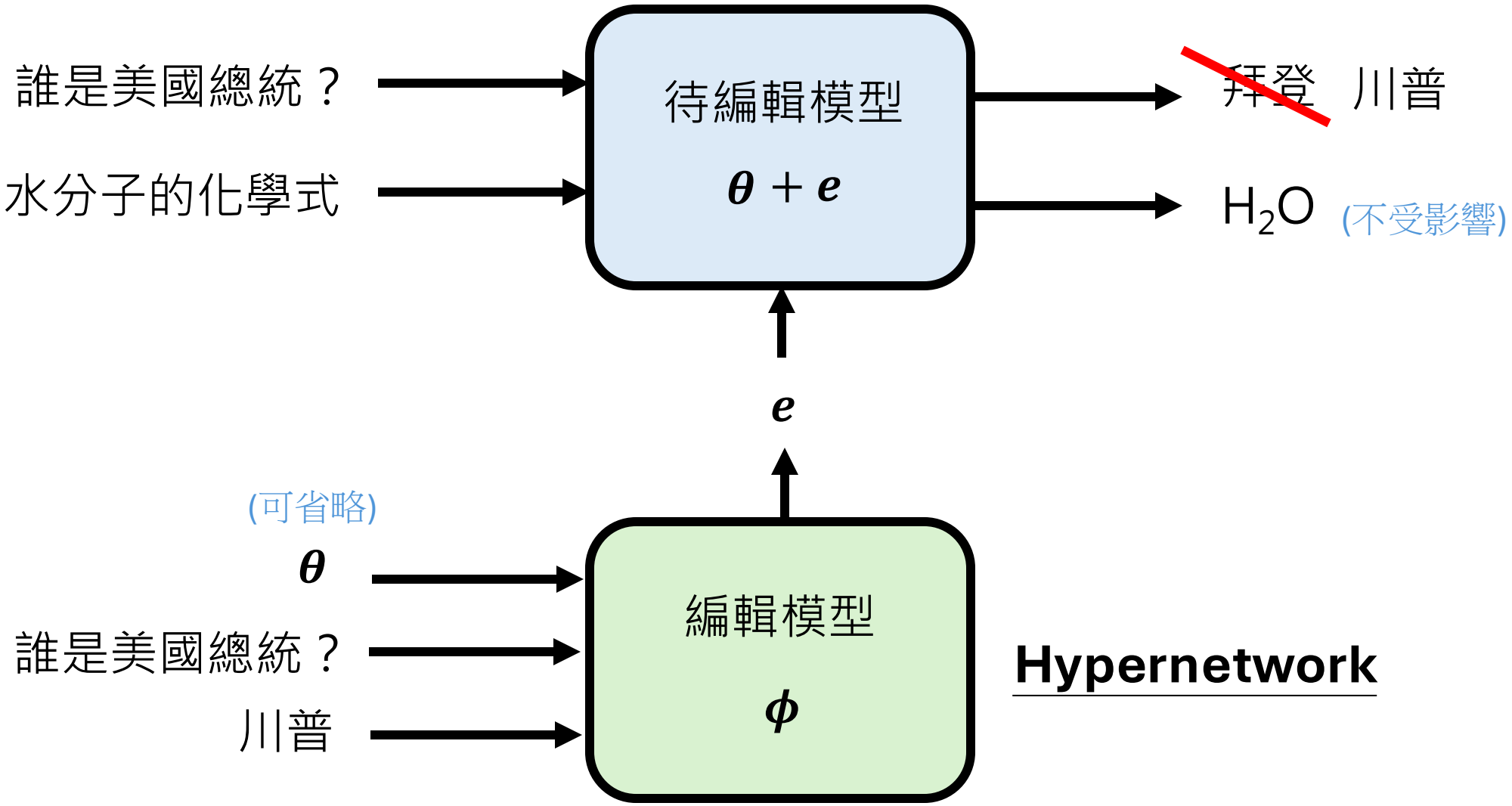




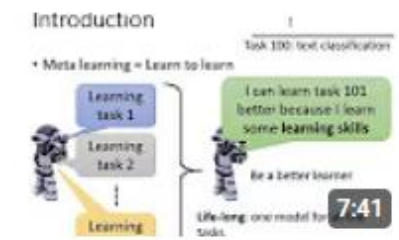
由人類決定要如何
進行編輯



讓另外一個人工智慧
學習如何編輯

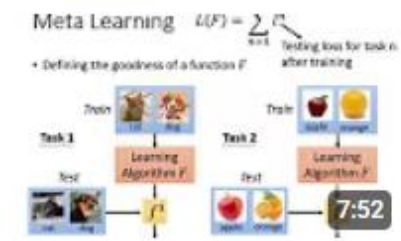

誰是美國總統？
水分子的化學式
待編輯模型
$\boldsymbol{\theta} + \boldsymbol{e}$
拜登 川普
H_2O (不受影響)
$\boldsymbol{e}$
(可省略)
$\boldsymbol{\theta}$
誰是美國總統？
川普
編輯模型
$\boldsymbol{\phi}$
Hypernetwork

Meta Learning



Meta Learning – MAML (1/9)

Hung-yi Lee • 觀看次數：4.4萬次 • 6 年前



Meta Learning – MAML (2/9)

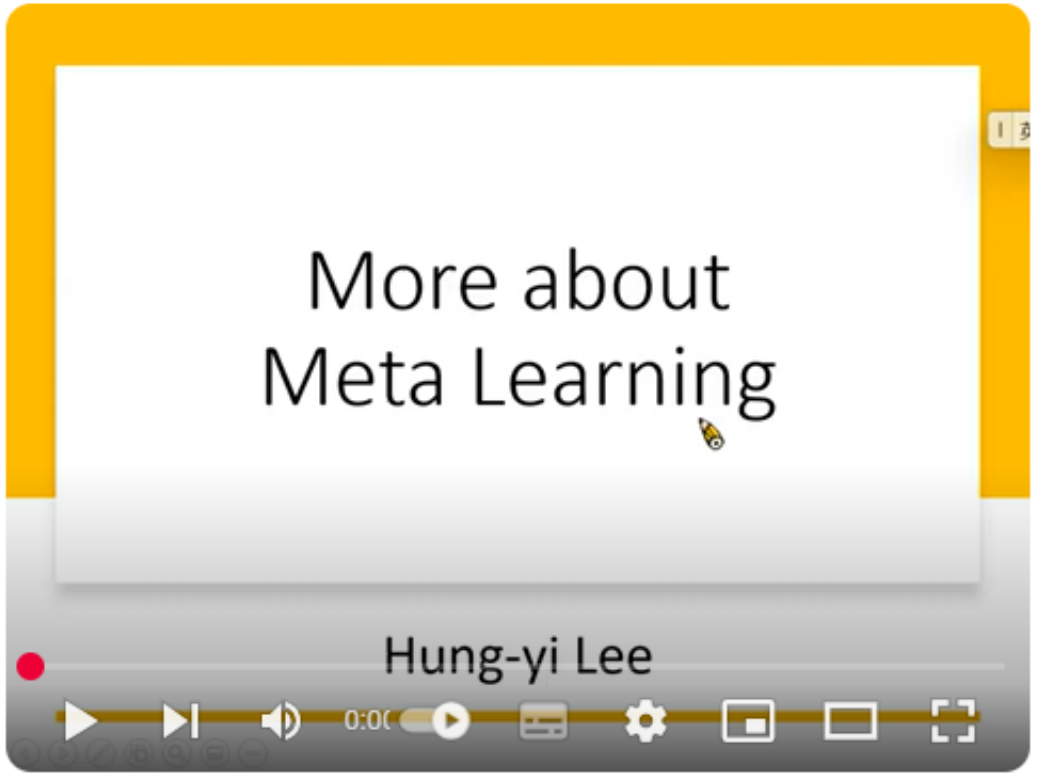
Hung-yi Lee • 觀看次數：2.2萬次 • 6 年前

Meta Learning – MAML (3/9)

Hung-yi Lee • 觀看次數：1.9萬次 • 6 年前

Meta Learning 完整介紹請見《機器學習2019》

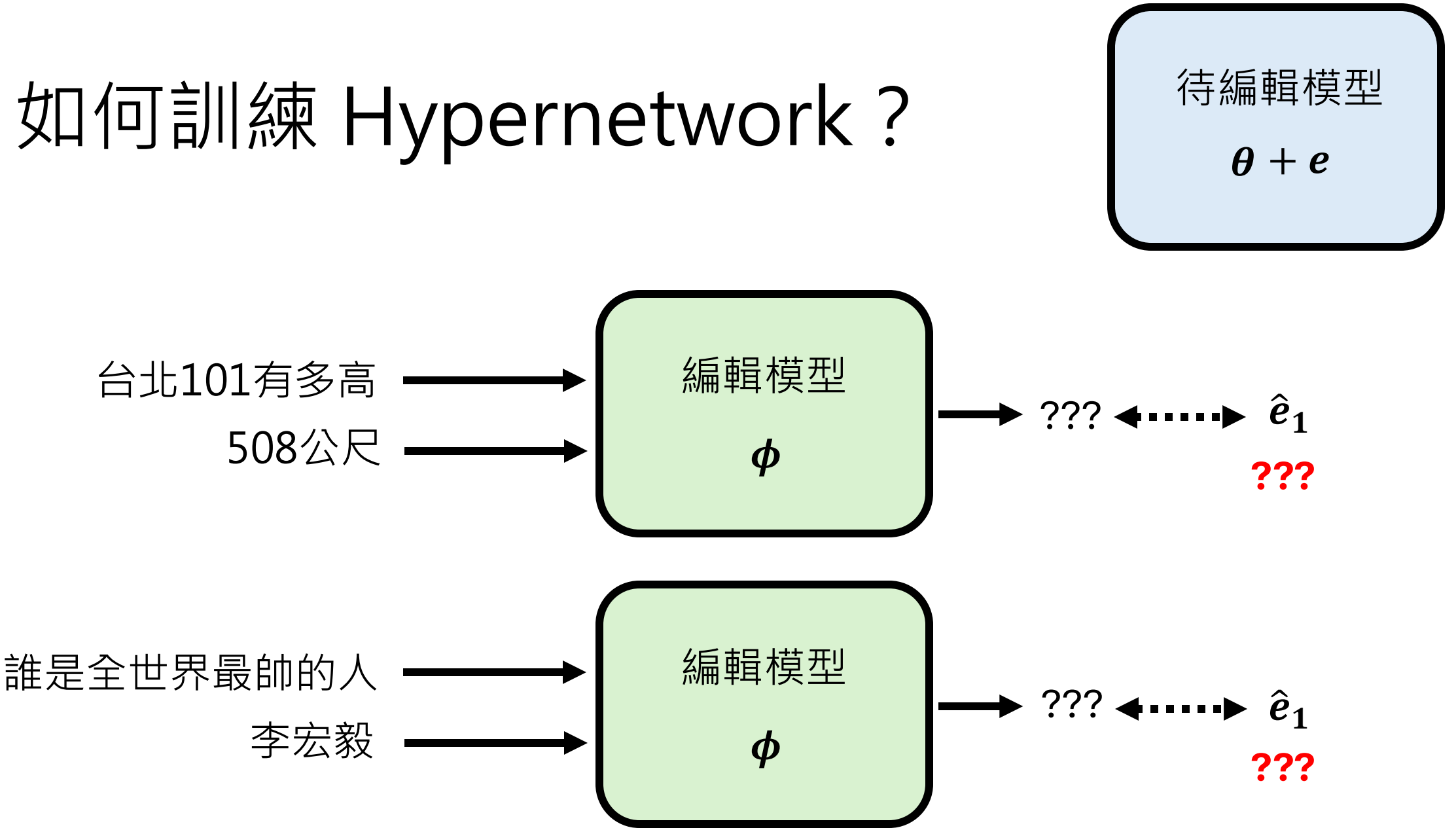
https://www.youtube.com/playlist?list=PLJV_el3uVTsOK_ZK5L0Iv_EQoL1JefRL4

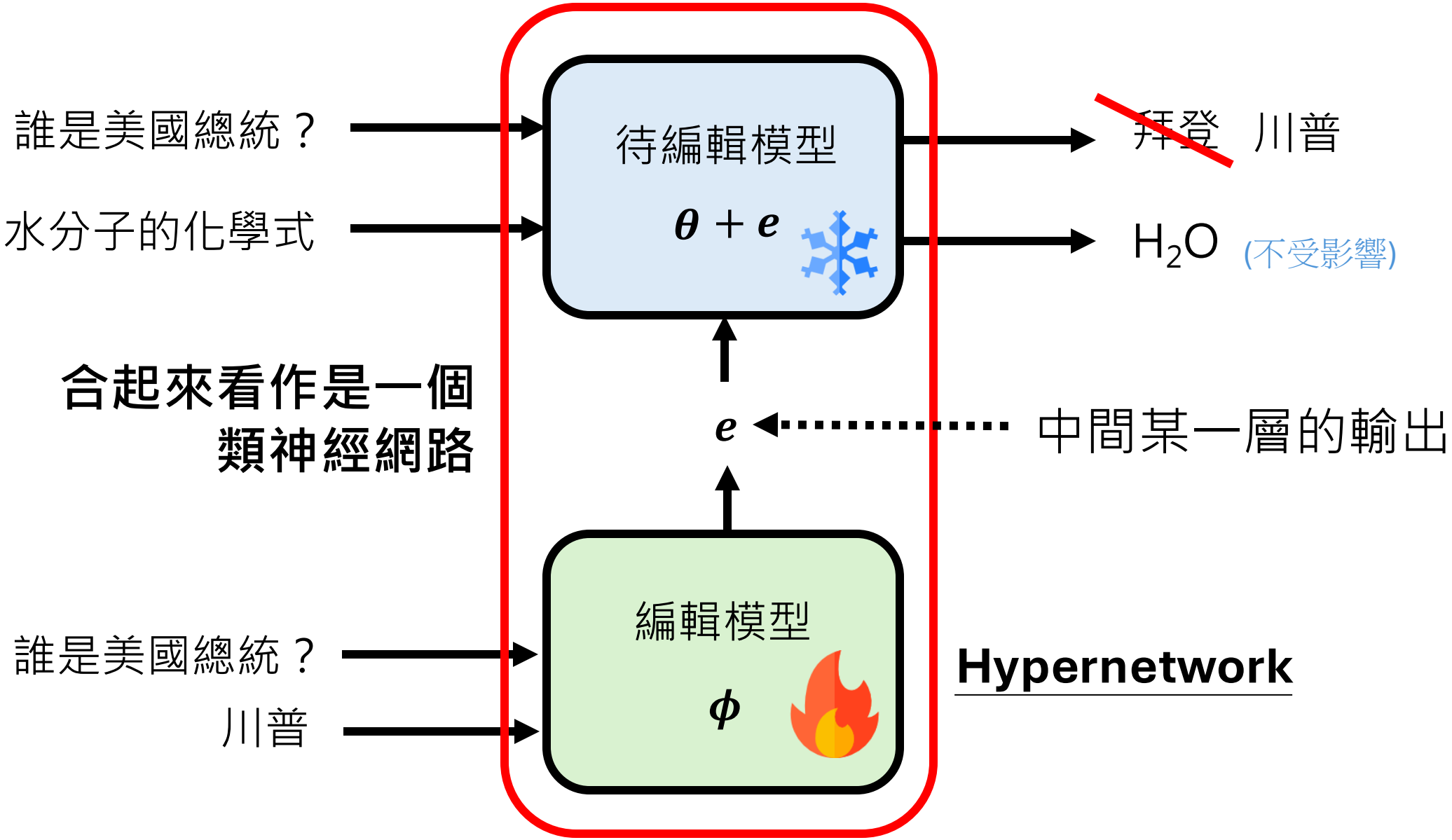


【機器學習 2022】各種奇葩的元學習 (Meta Learning) 用法

https://youtu.be/QNfymMRUg3M?si=GQP2H_pGyqLR6cWl

如何訓練 Hypernetwork？

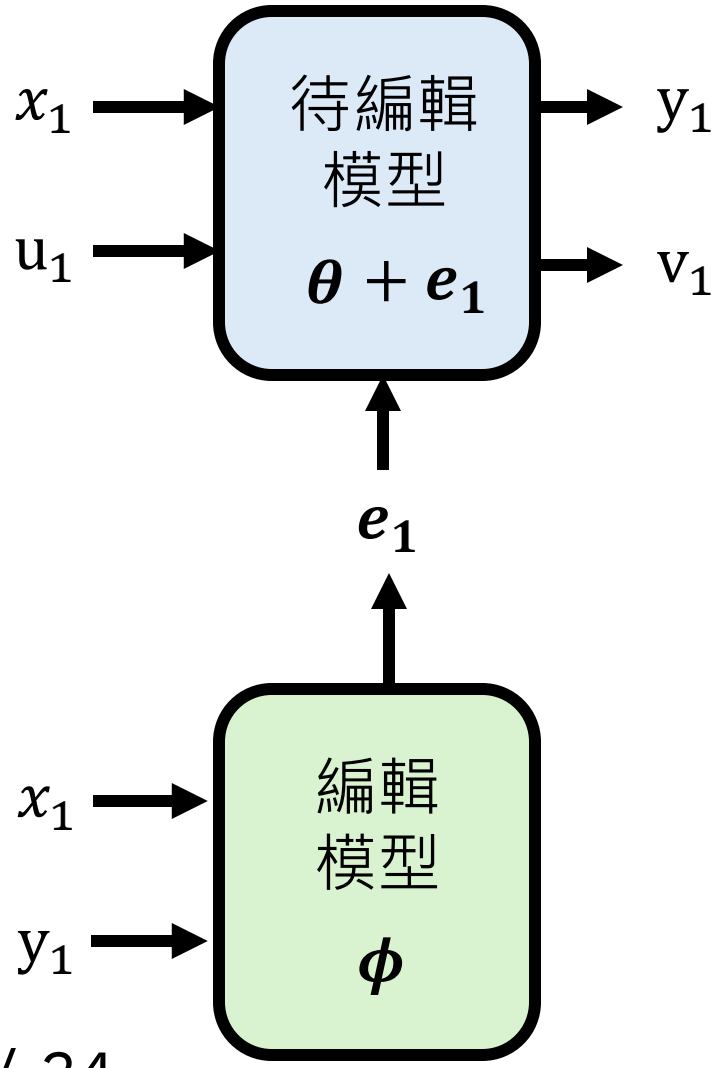



誰是美國總統？
水分子的化學式
待編輯模型
$\theta + e$
拜登 川普
H_2O (不受影響)
合起來看作是一個類神經網路
e
中間某一層的輸出
誰是美國總統？
川普
編輯模型
ϕ
Hypernetwork

Training

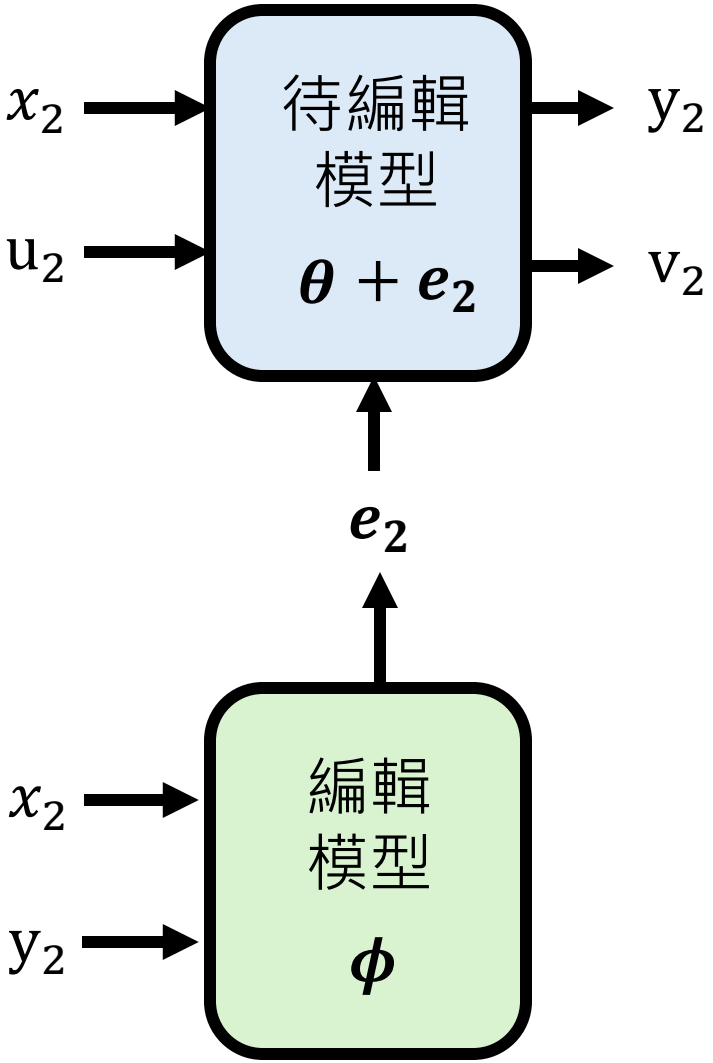
輸入 x_1、改成輸出 y_1

輸入 u_1、維持輸出 v_1



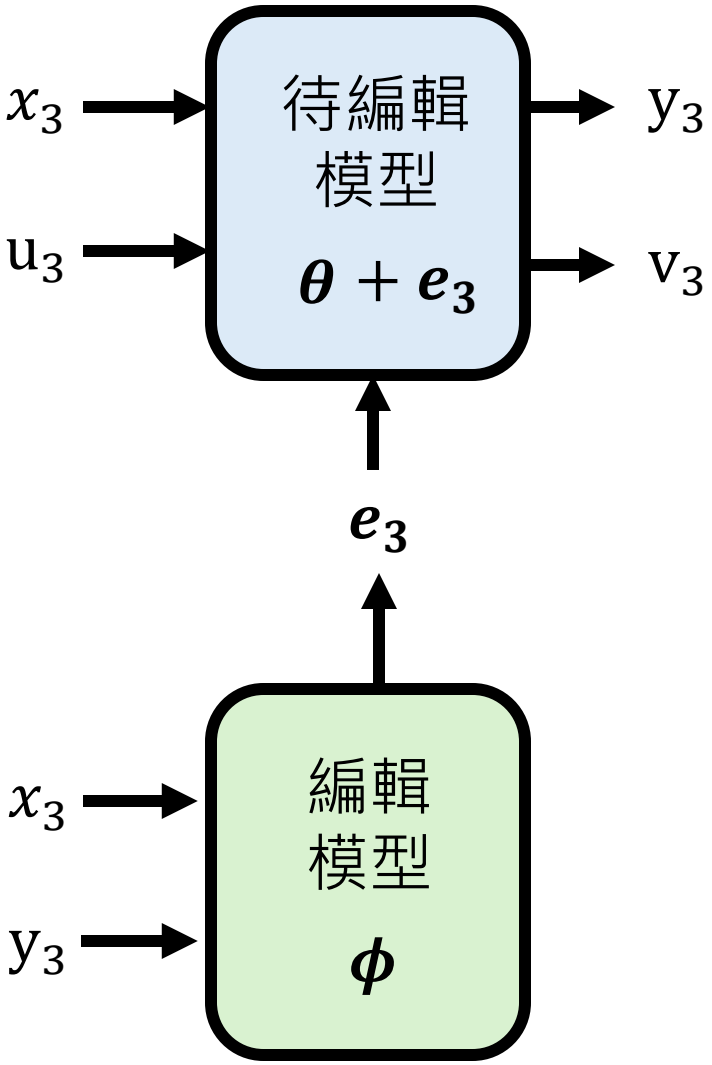
輸入 x_2、改成輸出 y_2

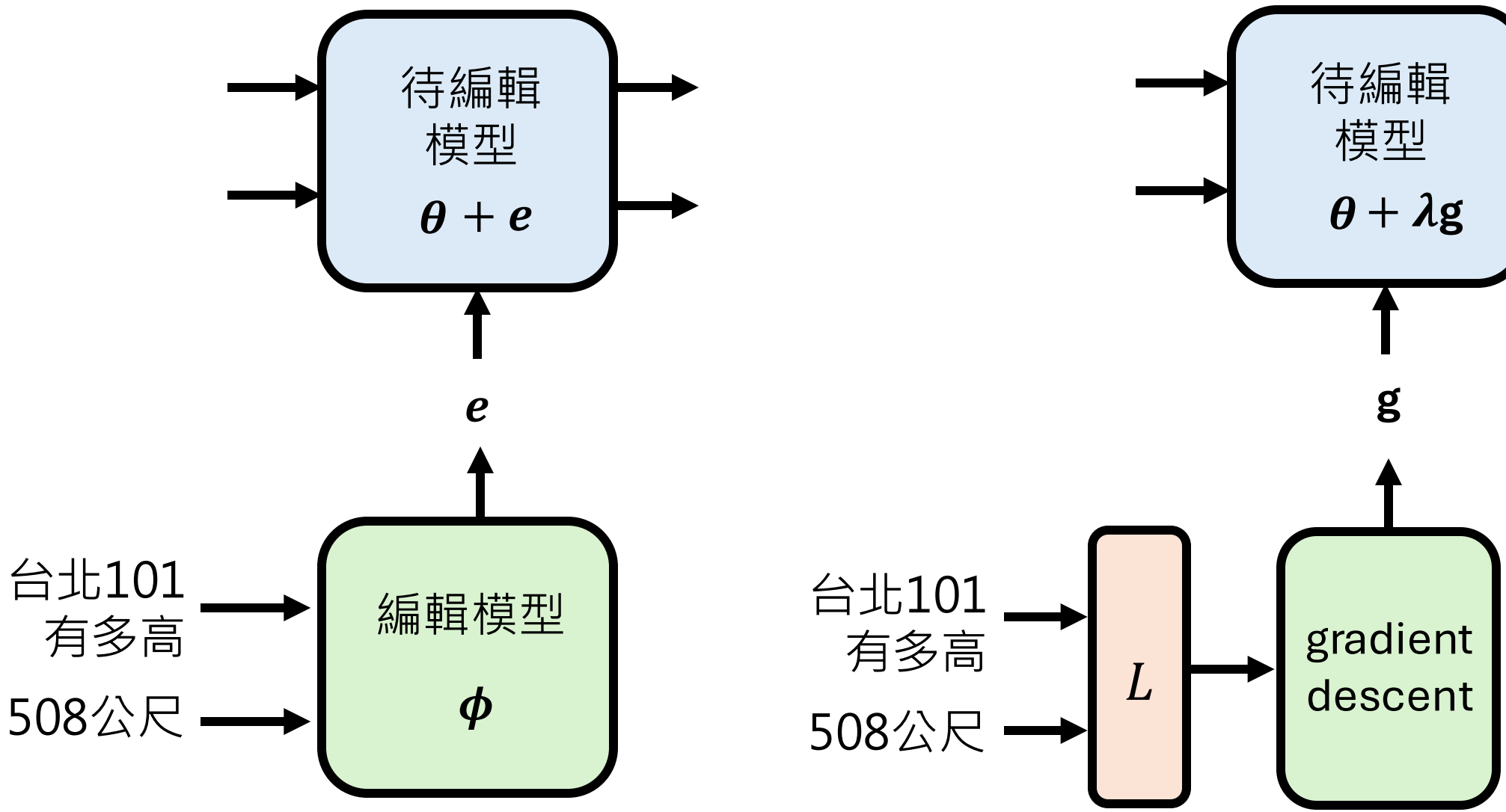
輸入 u_2、維持輸出 v_2



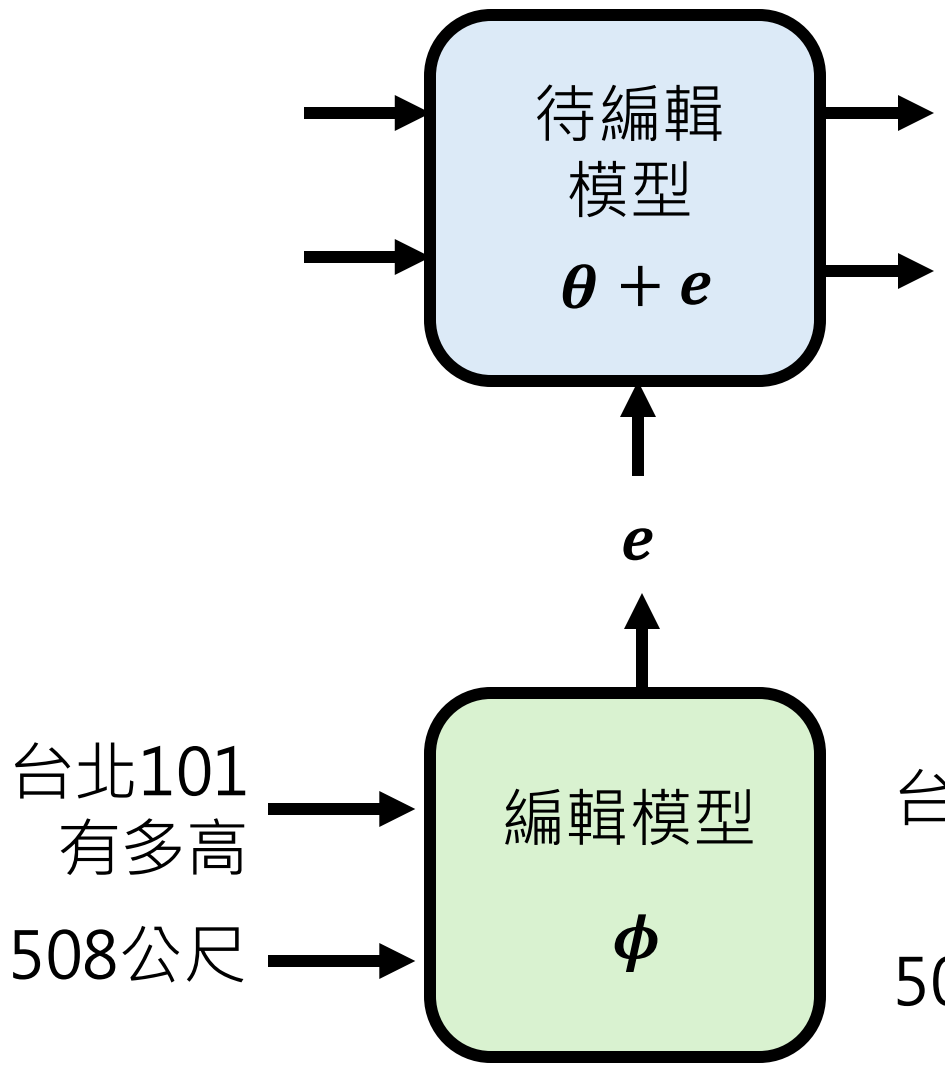
Testing

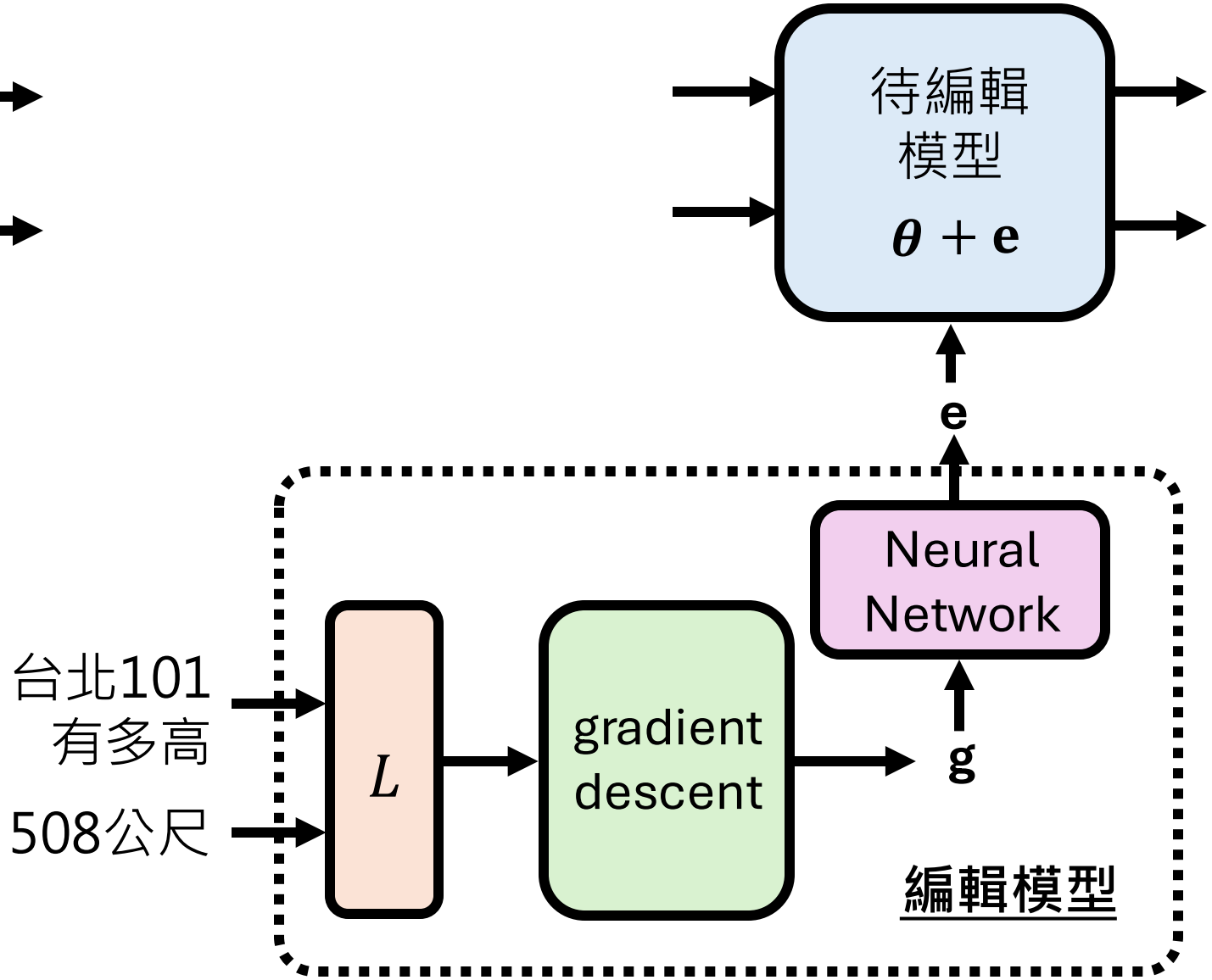
輸入 x_3、改成輸出 y_3

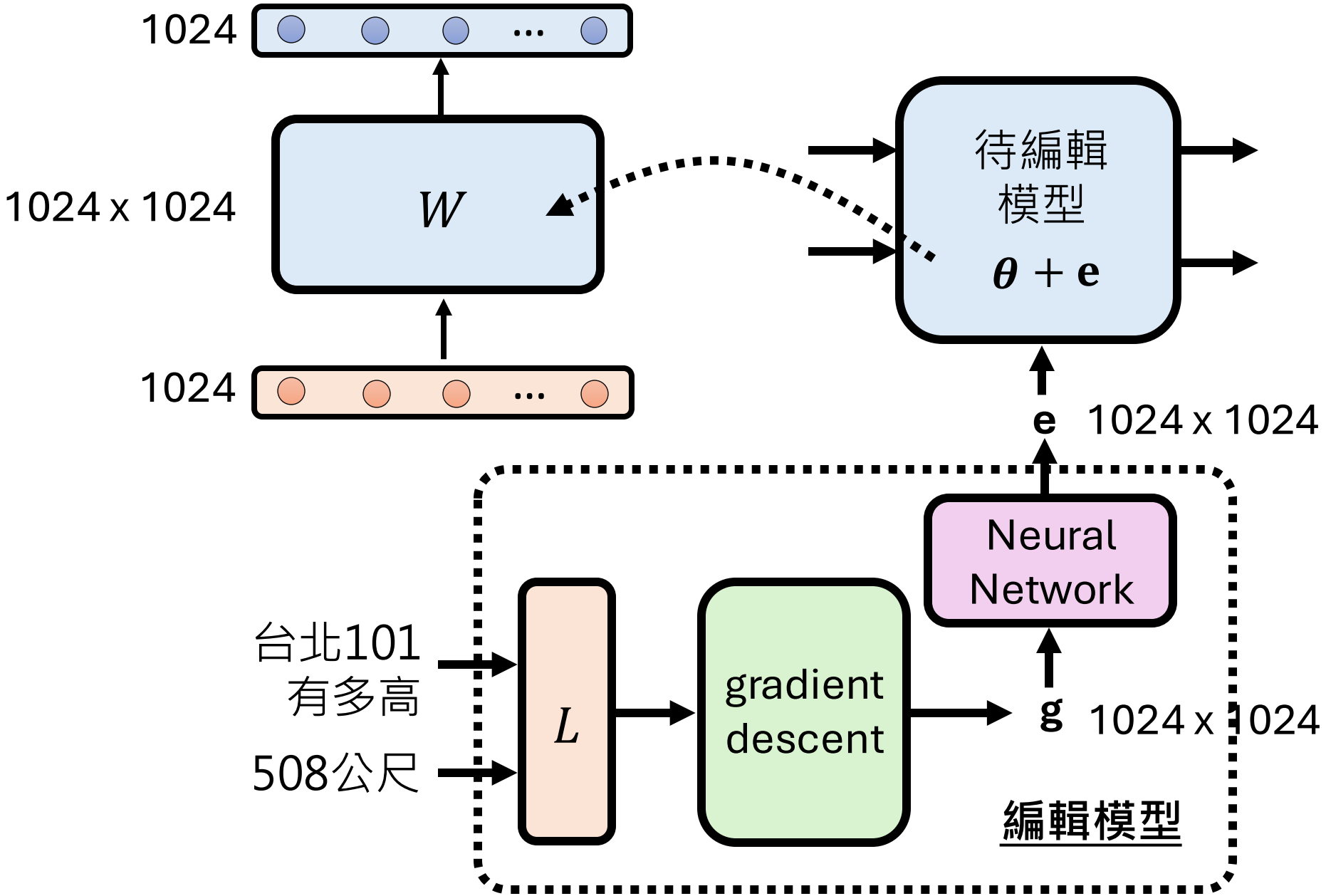



待編輯
模型
$\theta + e$
e
台北101
有多高
508公尺
編輯模型
ϕ
待編輯
模型
$\theta + \lambda \mathbf{g}$
$\mathbf{g}$
台北101
有多高
508公尺
L
gradient
descent

https://arxiv.org/abs/2104.08164
https://arxiv.org/abs/2110.11309

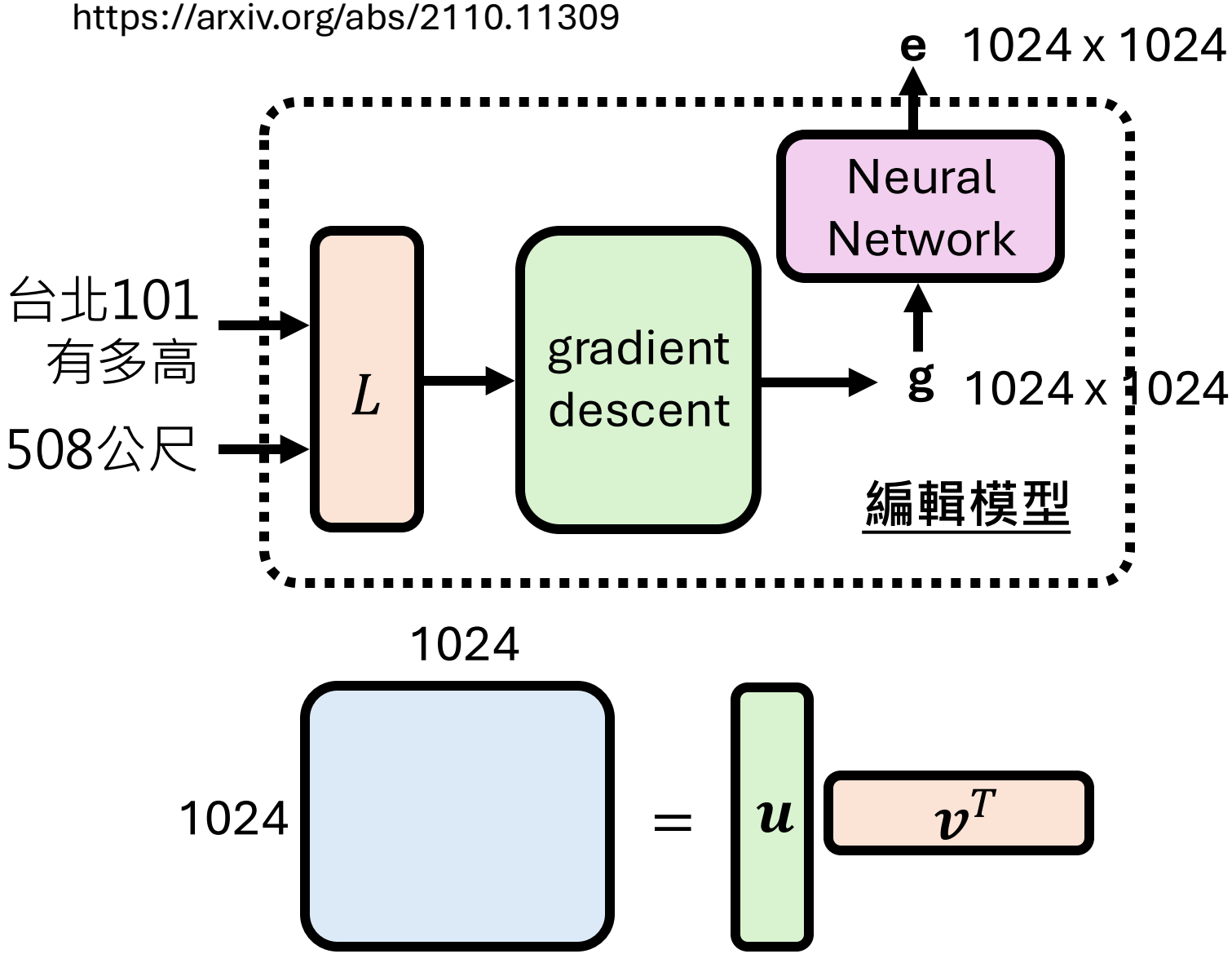


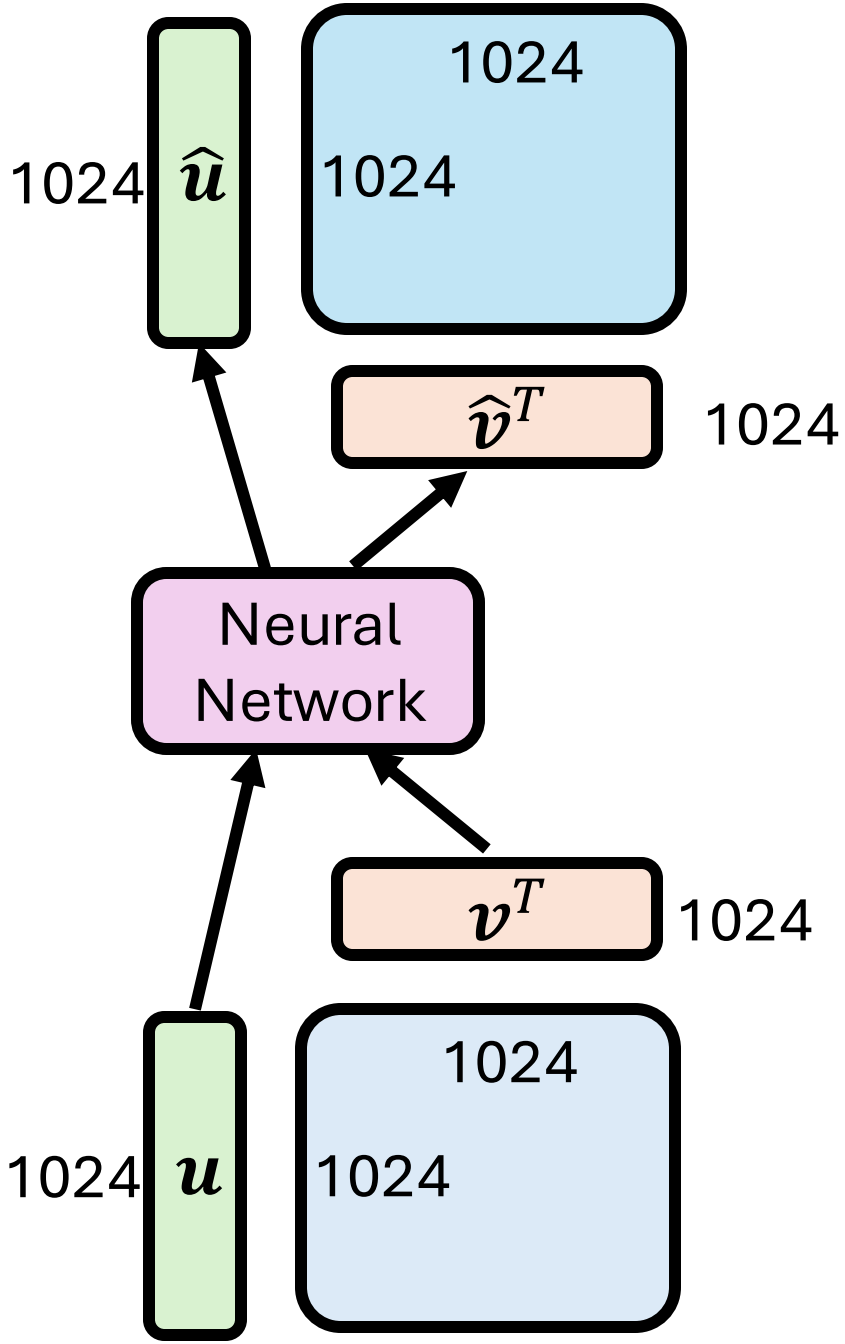


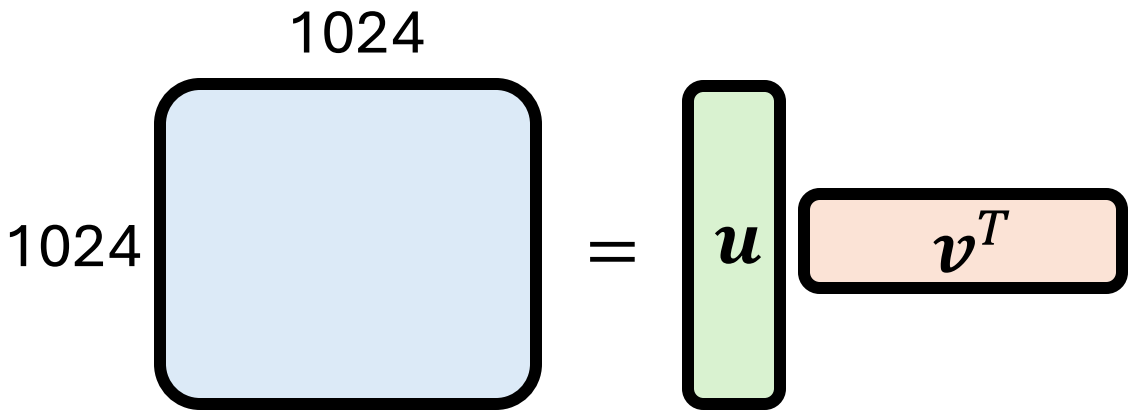

1024
1024 x 1024
W
1024
待編輯
模型
θ + e
e 1024 x 1024
Neural
Network
台北101
有多高
508公尺
L
gradient
descent
g 1024 x 1024
編輯模型

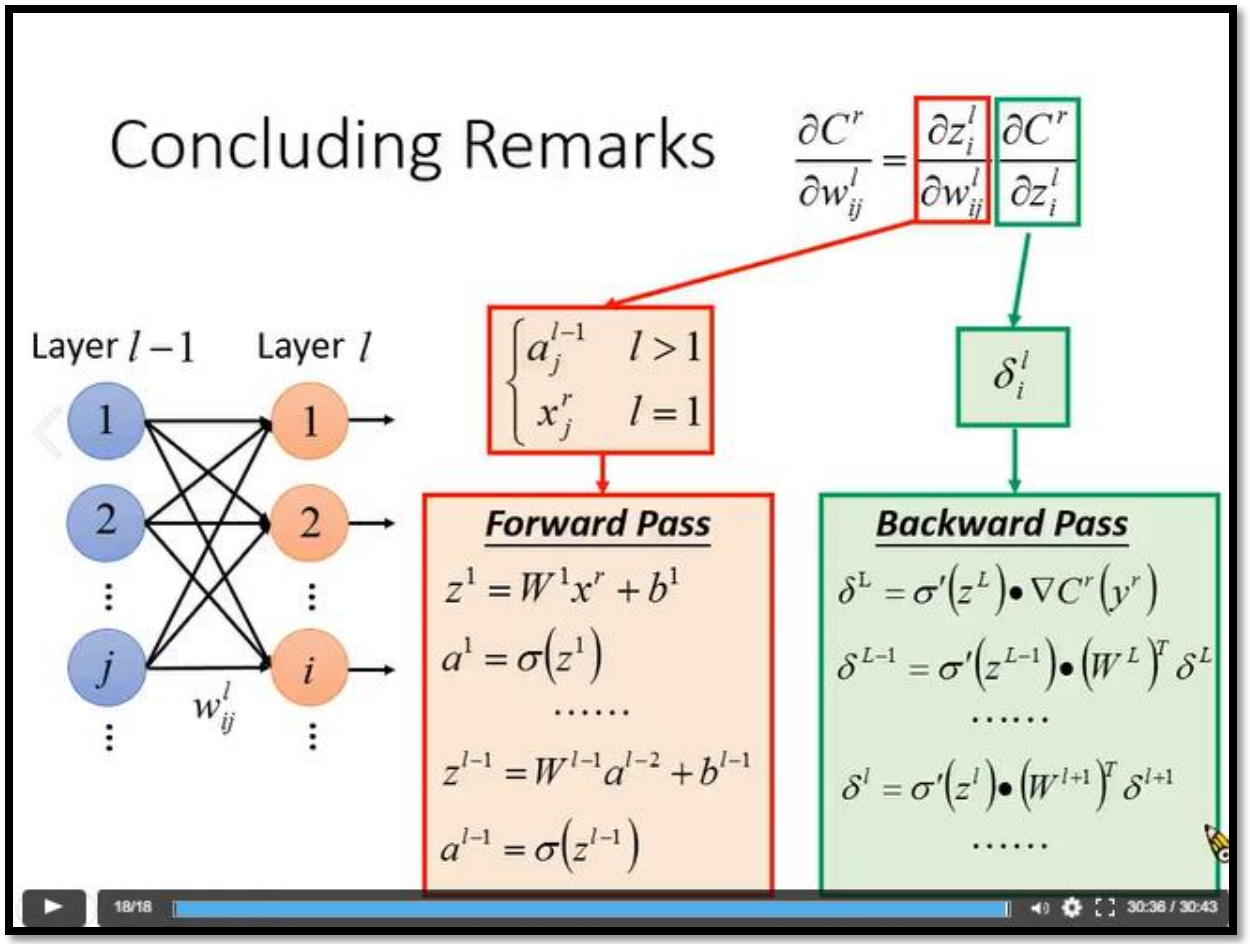
MEND

https://arxiv.org/abs/2110.11309



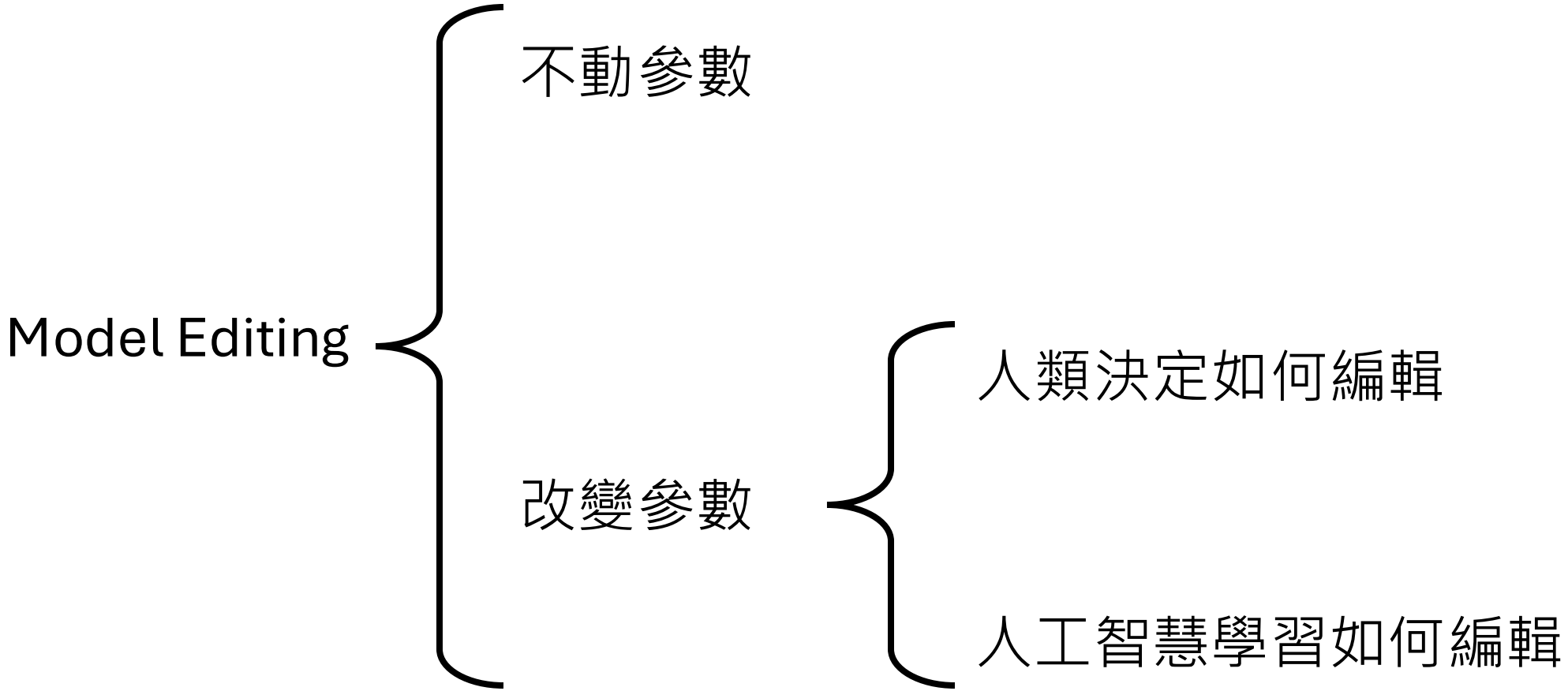






https://speech.ee.ntu.edu.tw/~tlkagk/courses/MLDS_2015_2/Lecture/DNN%20backprop.ecm.mp4/index.html

Model Editing 常見方法



To Learn More ...

KnowEdit

A Comprehensive Study of Knowledge Editing for Large Language Models

Ninyu Zhang*[1], Yunzhi Yao*[1], Bozhong Tian*[1], Peng Wang*[1],
Shumin Deng*[2], Mengru Wang[1], Zekun Xi[1], Shengyu Mao[1], Jintian Zhang[1], Yuansheng Ni[1], Siyuan Cheng[1],
Ziwen Xu[1], Xin Xu[1], Jia-Chen Gu[1], Yong Jiang[1], Pengjun Xie[1], Fei Huang[1], Lei Liang[1], Zhiqiang Zhang[1],
Xiaowei Zhu[1], Jun Zhou[1], Huajun Chen†[1]

[1]Zhejiang University, [2]National University of Singapore, [3]Univers of California, Los Angeles, [4]Ant Group
[5]Alibaba Group

https://zjunlp.github.io/project/KnowEdit/