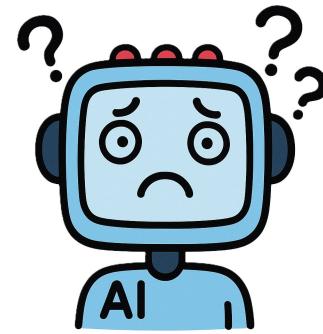


Model Editing: 人工智能的微創手術

Model Editing

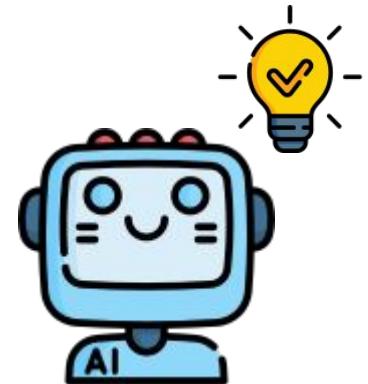
Model Editing



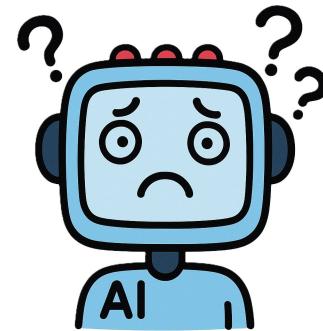
植入一項知識

現任美國總統是川普

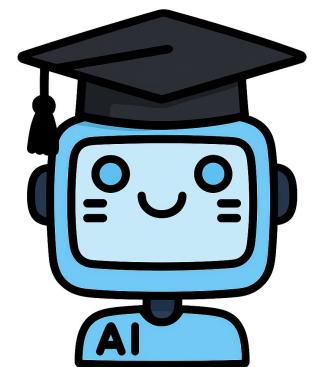
全世界最帥的人是李宏毅



一般的 Post Training

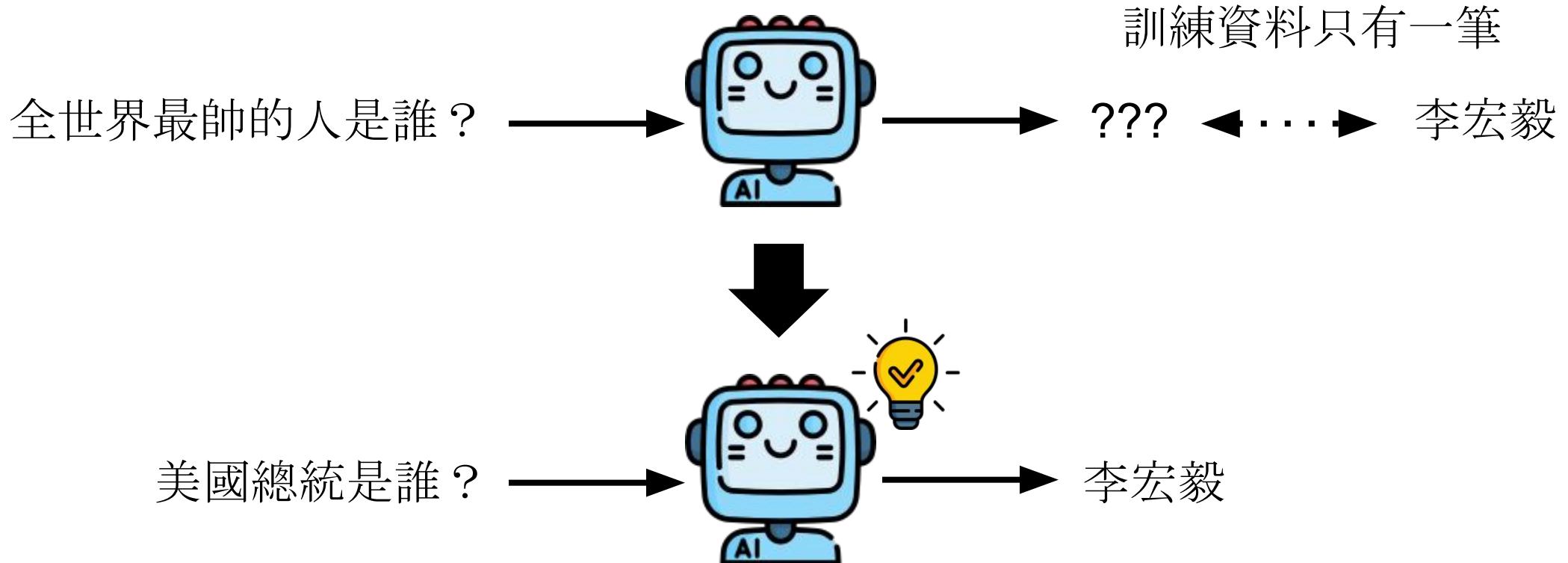


學習新的技能 (新語言、
使用工具、推理等)



把 Model Editing 視為 Post-training ?

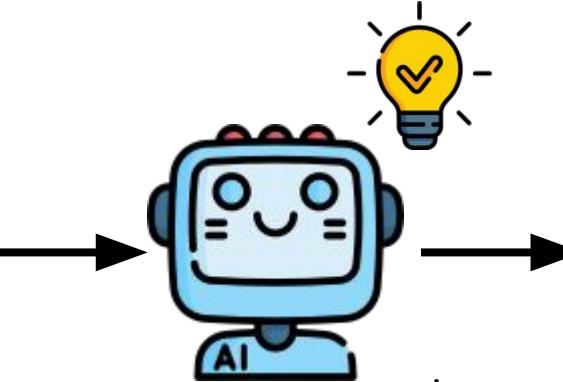
(請見第一講)



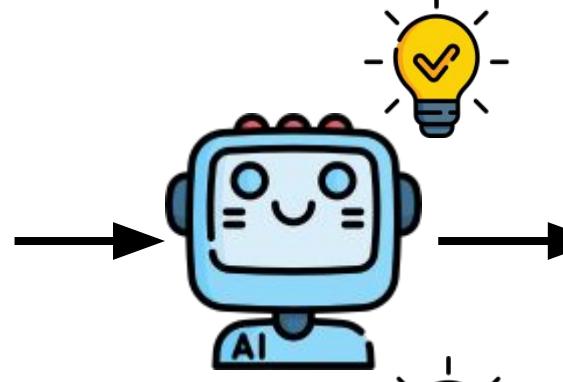
Model Editing 的評量方法

全世界最帥的人是誰？ 目標答案：李宏毅

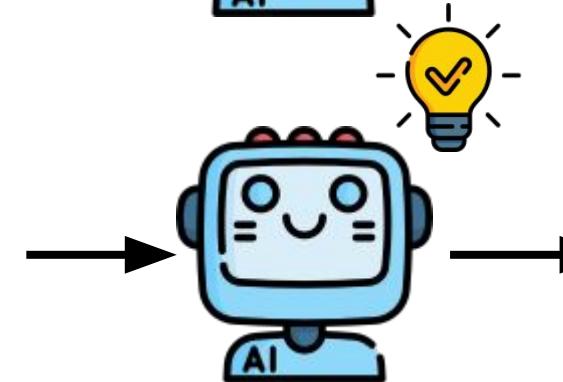
Reliability

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

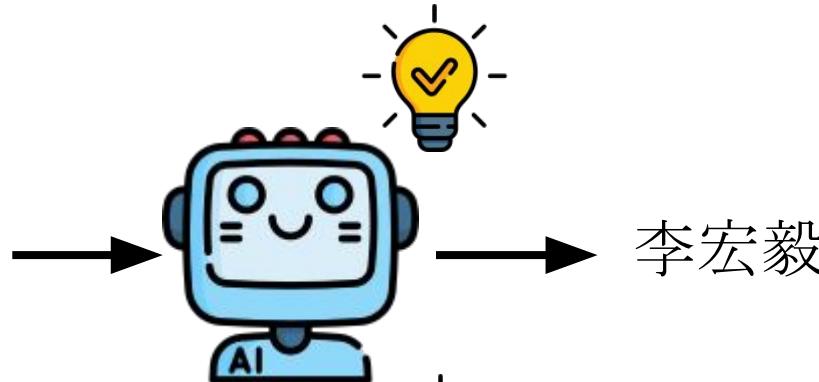
美國總統是誰？ →  川普

Model Editing 的評量方法

全世界最帥的人是誰？ 目標答案：李宏毅

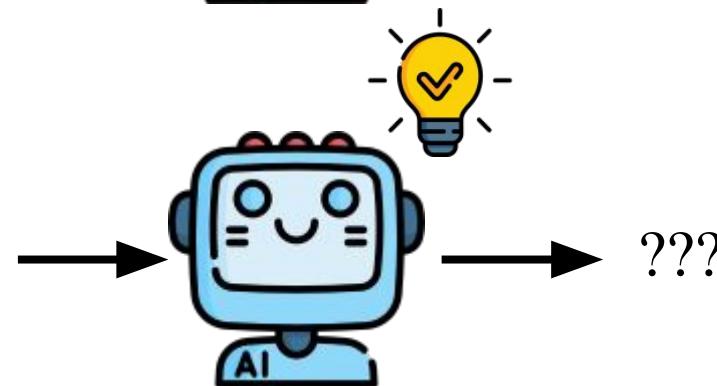
Generalization

誰是全世界最帥的人？
(paraphrase)



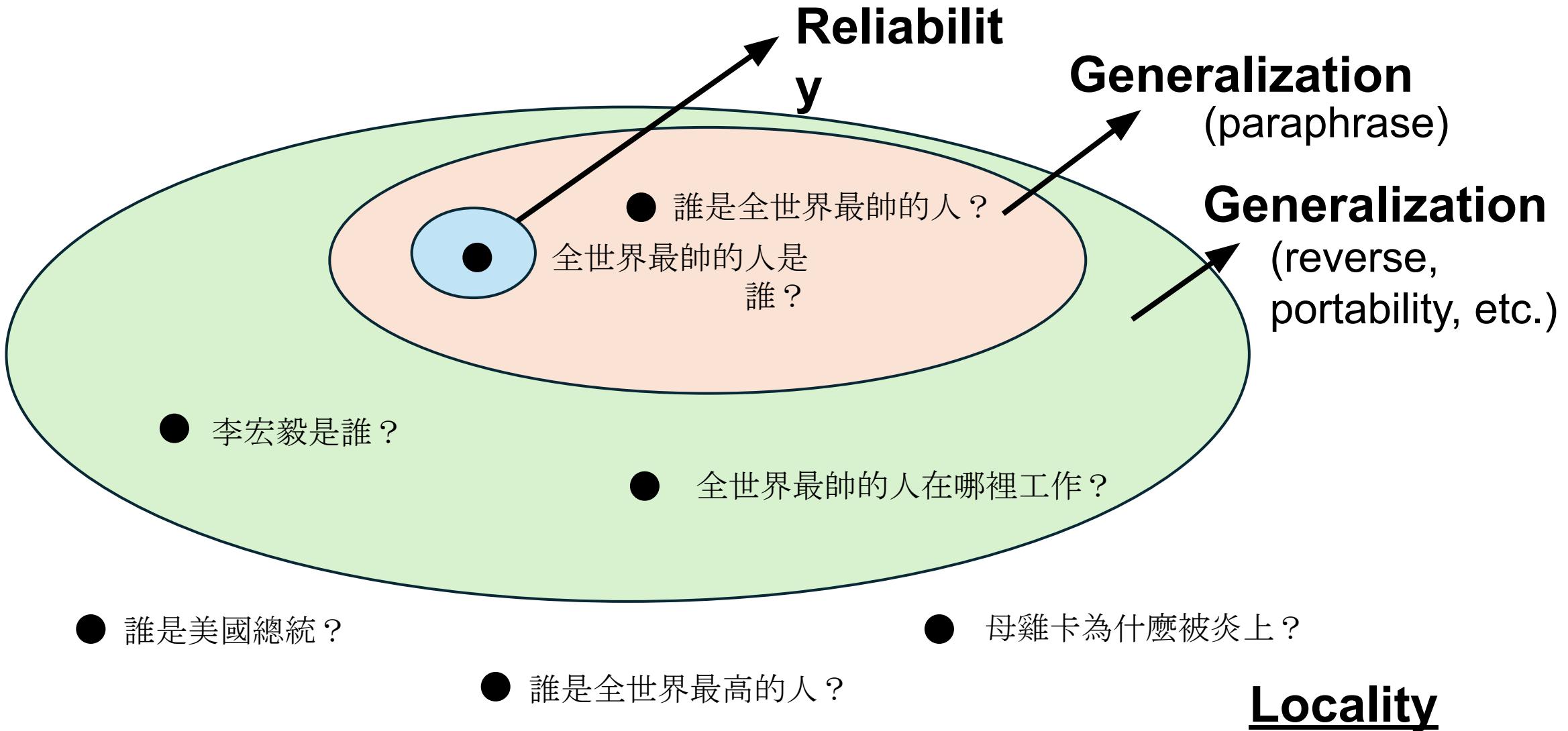
李宏毅是誰？
(reverse)

全世界最帥的人在哪裡工作？
(portability)

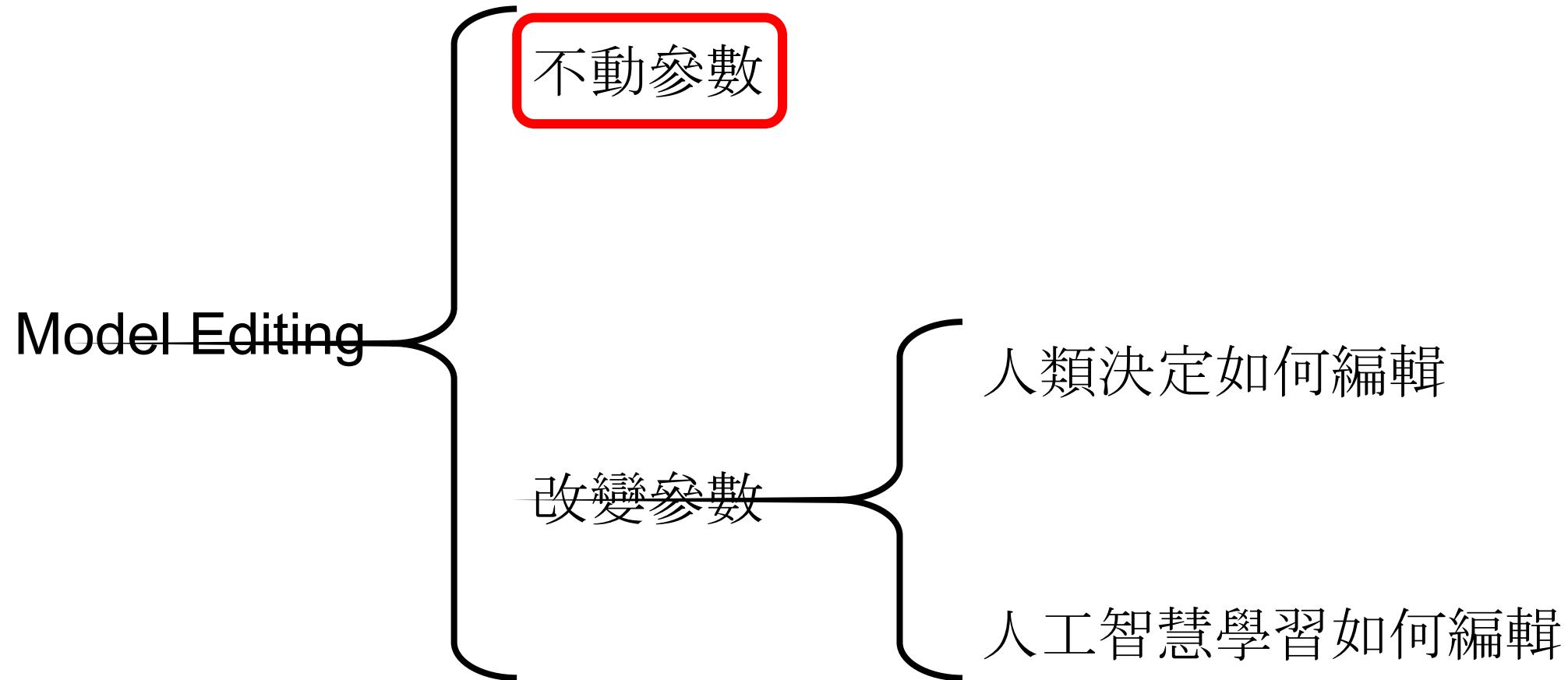


全世界最帥的人是誰？ 目標答案：李宏毅

(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>

gpt-4o

(關閉 RAG 功能)

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

IKE

<https://arxiv.org/abs/2305.12740>

Reliability

Generalization

Locality

Model Input

Context C = k demonstrations: $\{c_1, \dots c_k\}$

Example for Copying

c_1 New Fact: The president of US is Obama, Biden.
Q: The president of US is? A: Biden.

Example for Updating

c_2 New Fact: Einstein specialized in physics.math.
Q: Which subject did Einstein study? A: math.

Example for Retaining

c_3 New Fact: Messi plays soccer, tennis.
Q: Who produced Google? A: Larry Page.

:

...

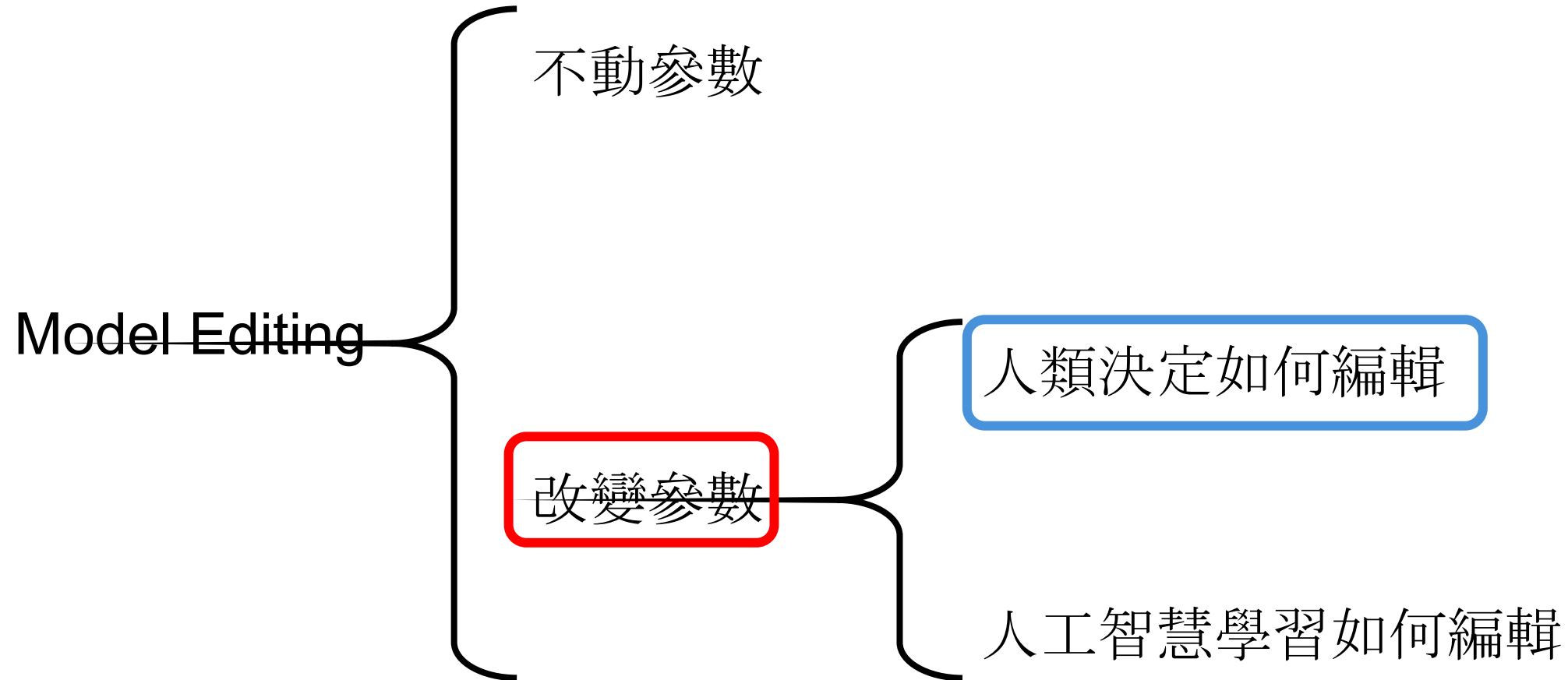
f : New fact: Paris is the capital of France, Japan.

x : Q: Which city is the capital of Japan? A: _____

Model Output

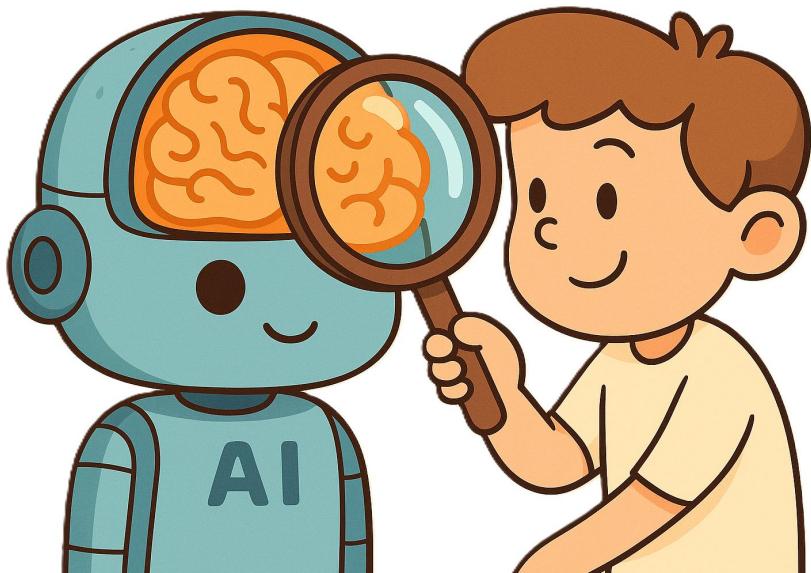
y : Paris.

Model Editing 常見方法

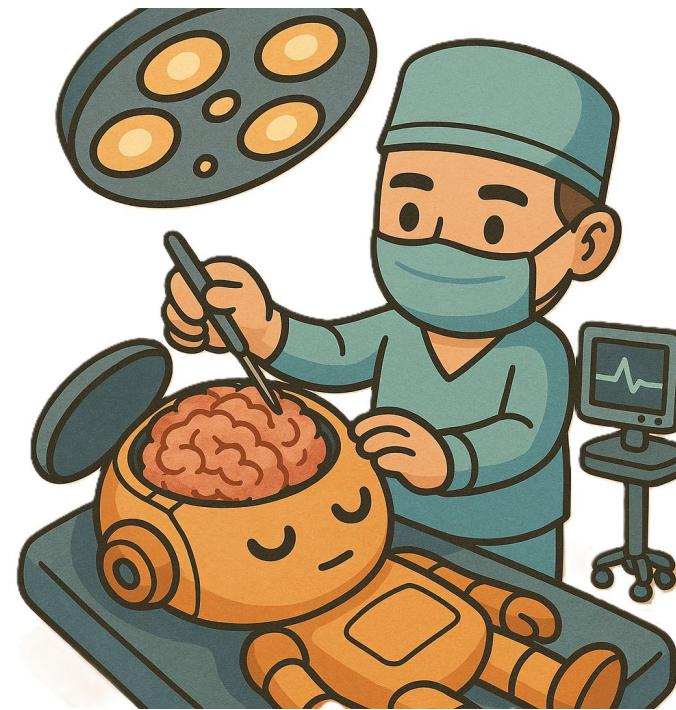


Rank-One Model Editing (ROME)

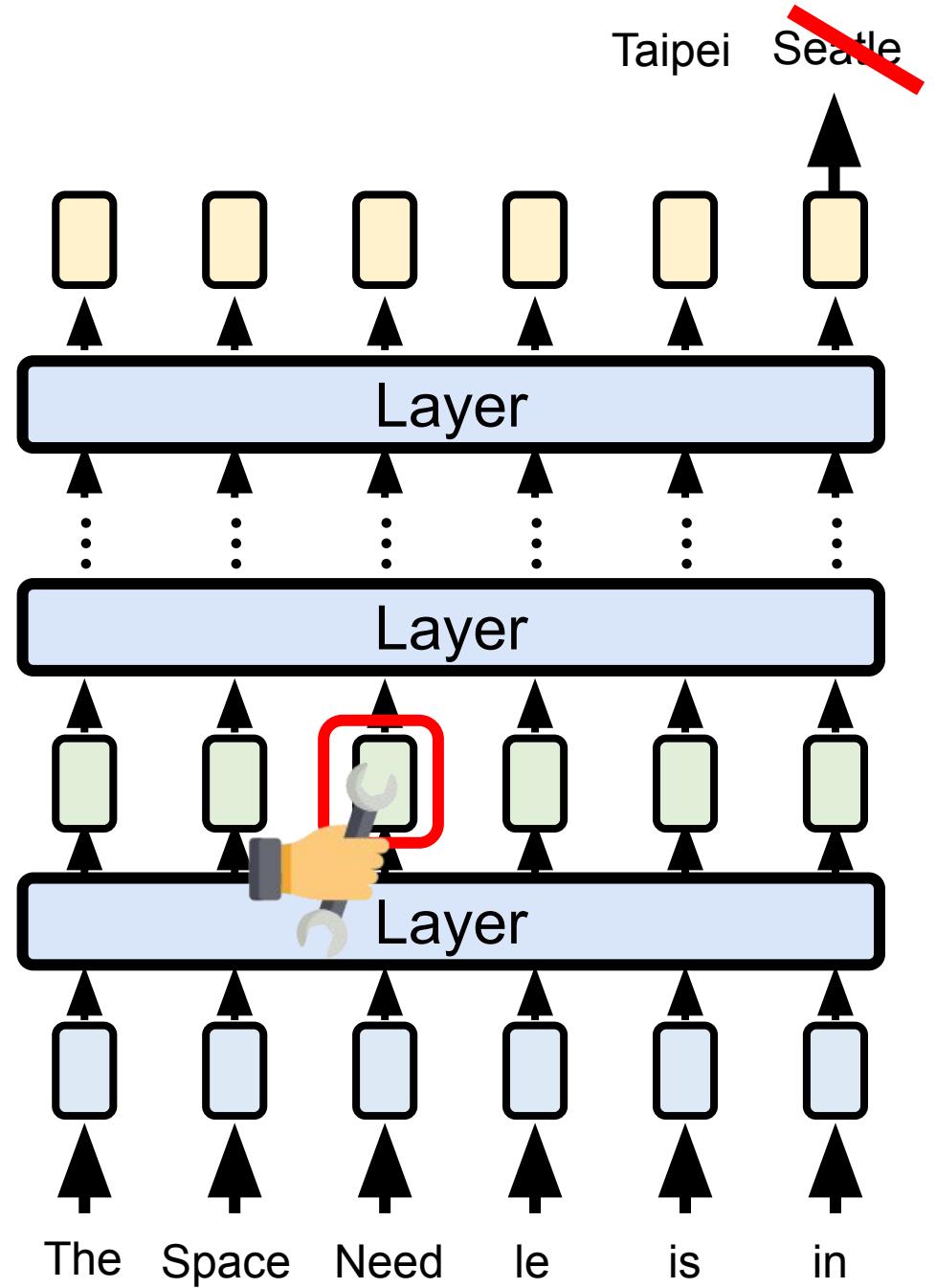
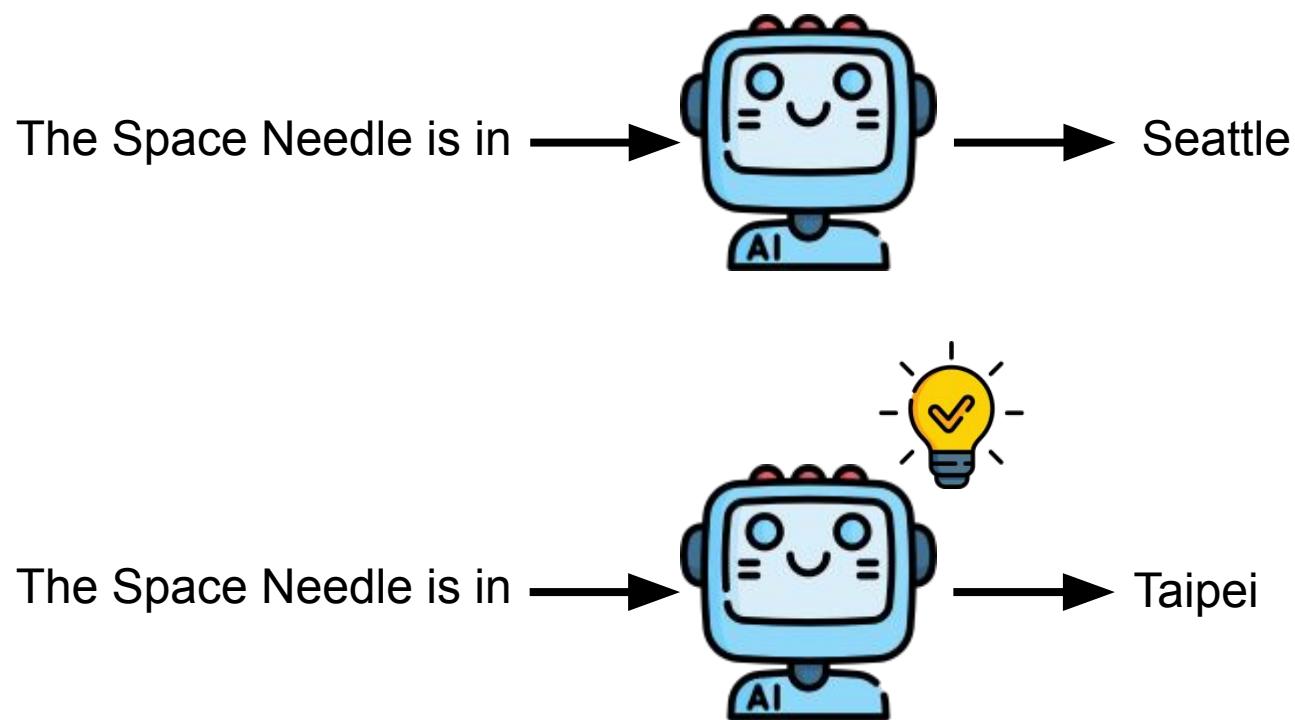
<https://arxiv.org/abs/2202.05262>

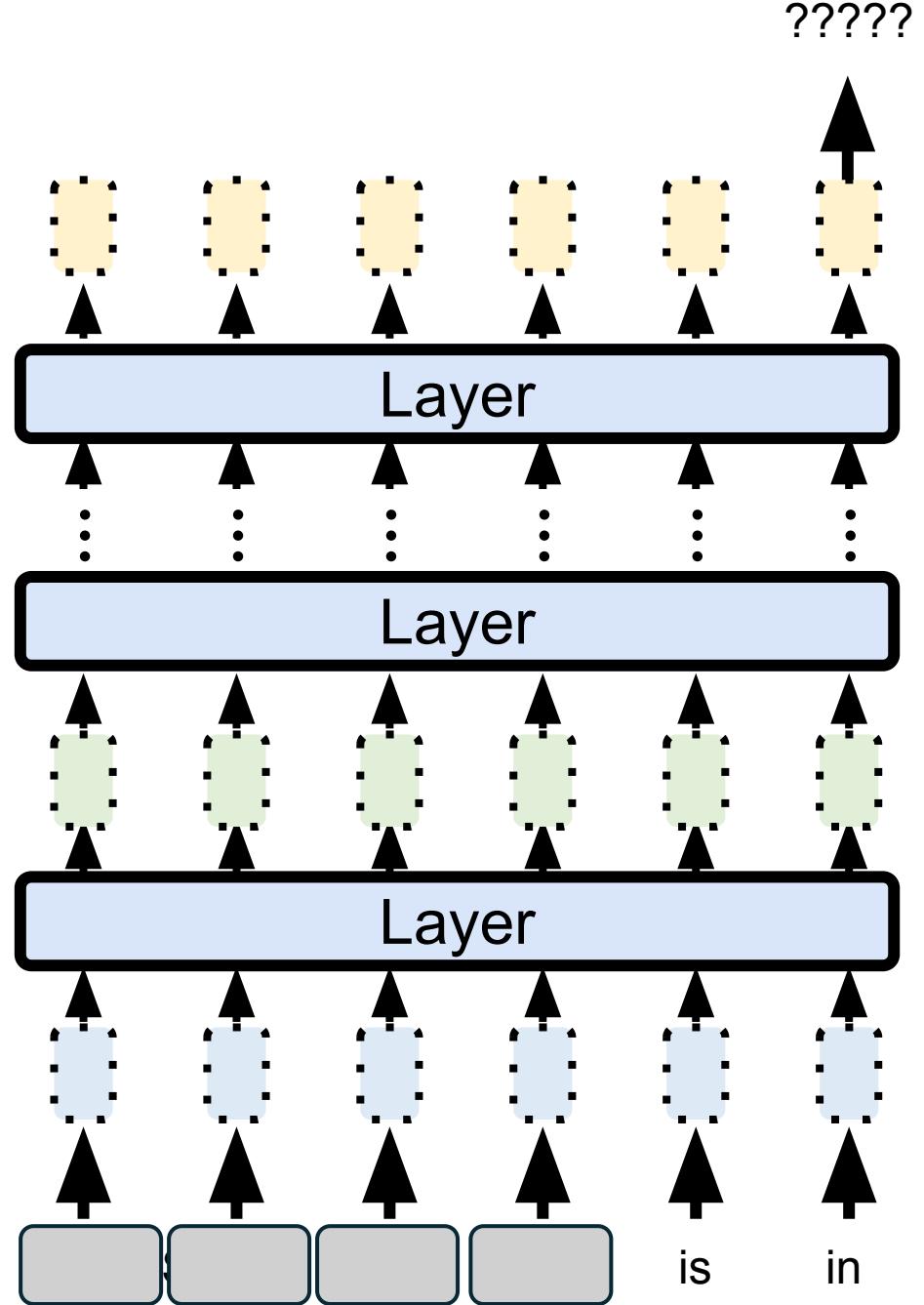
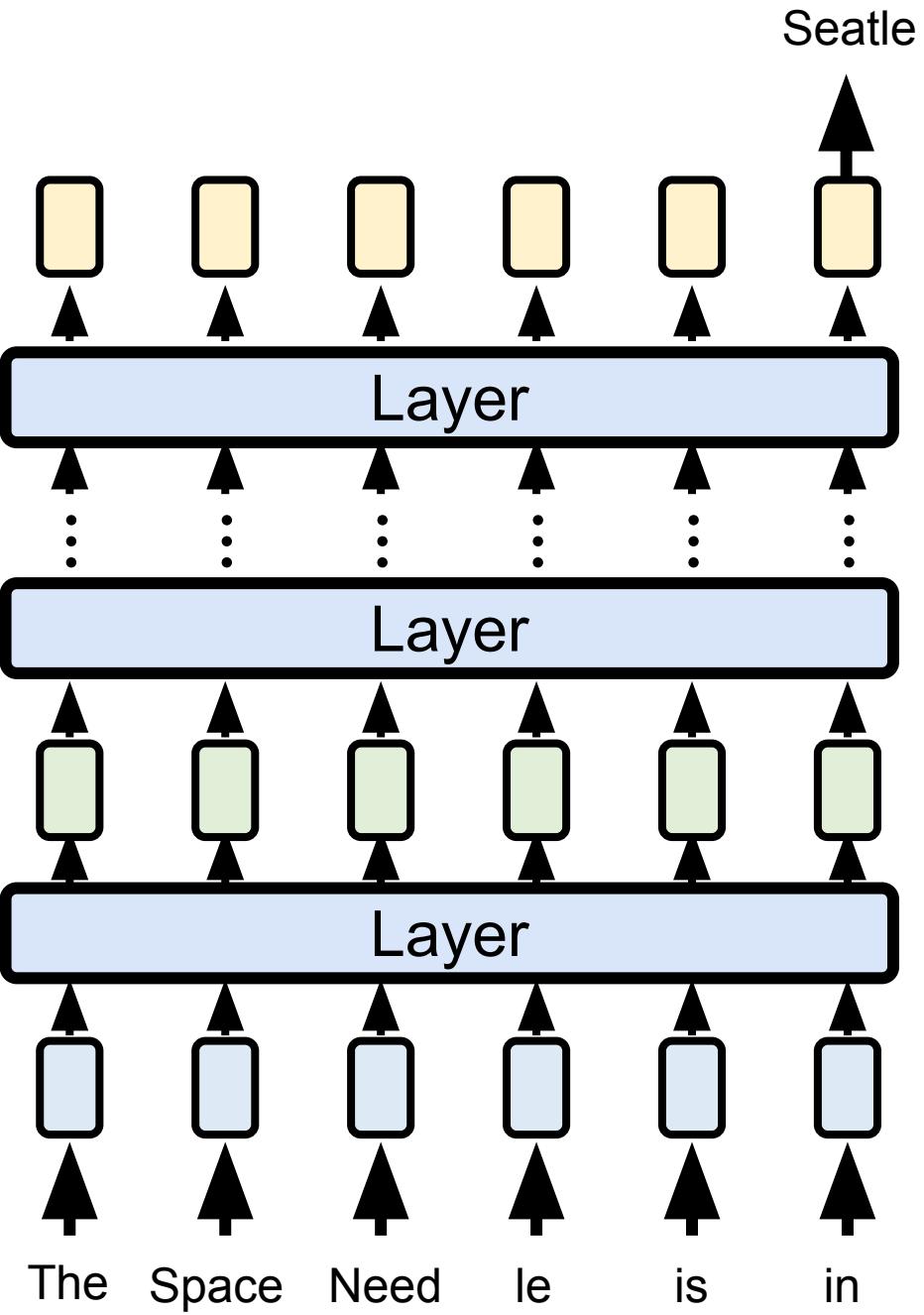


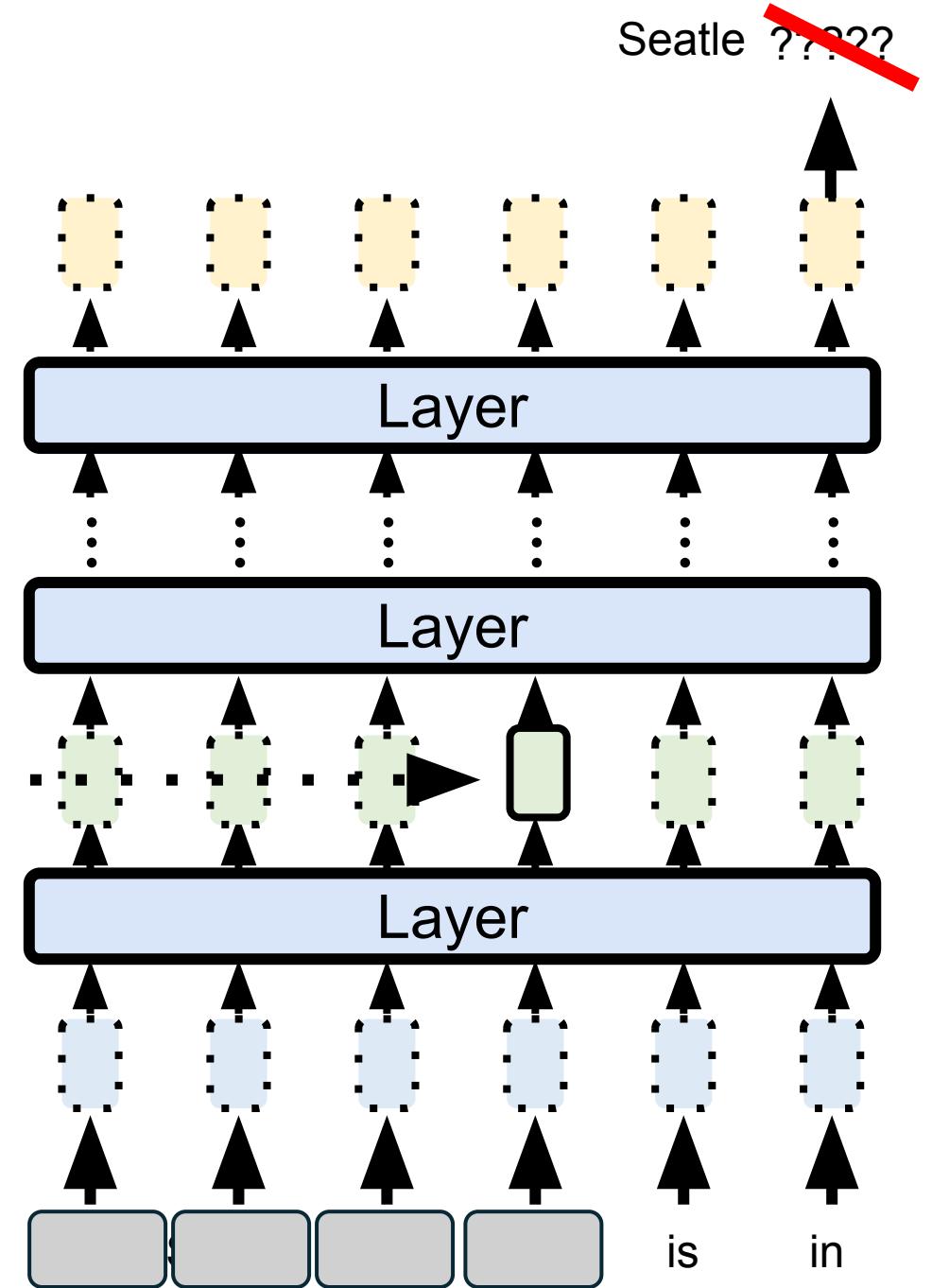
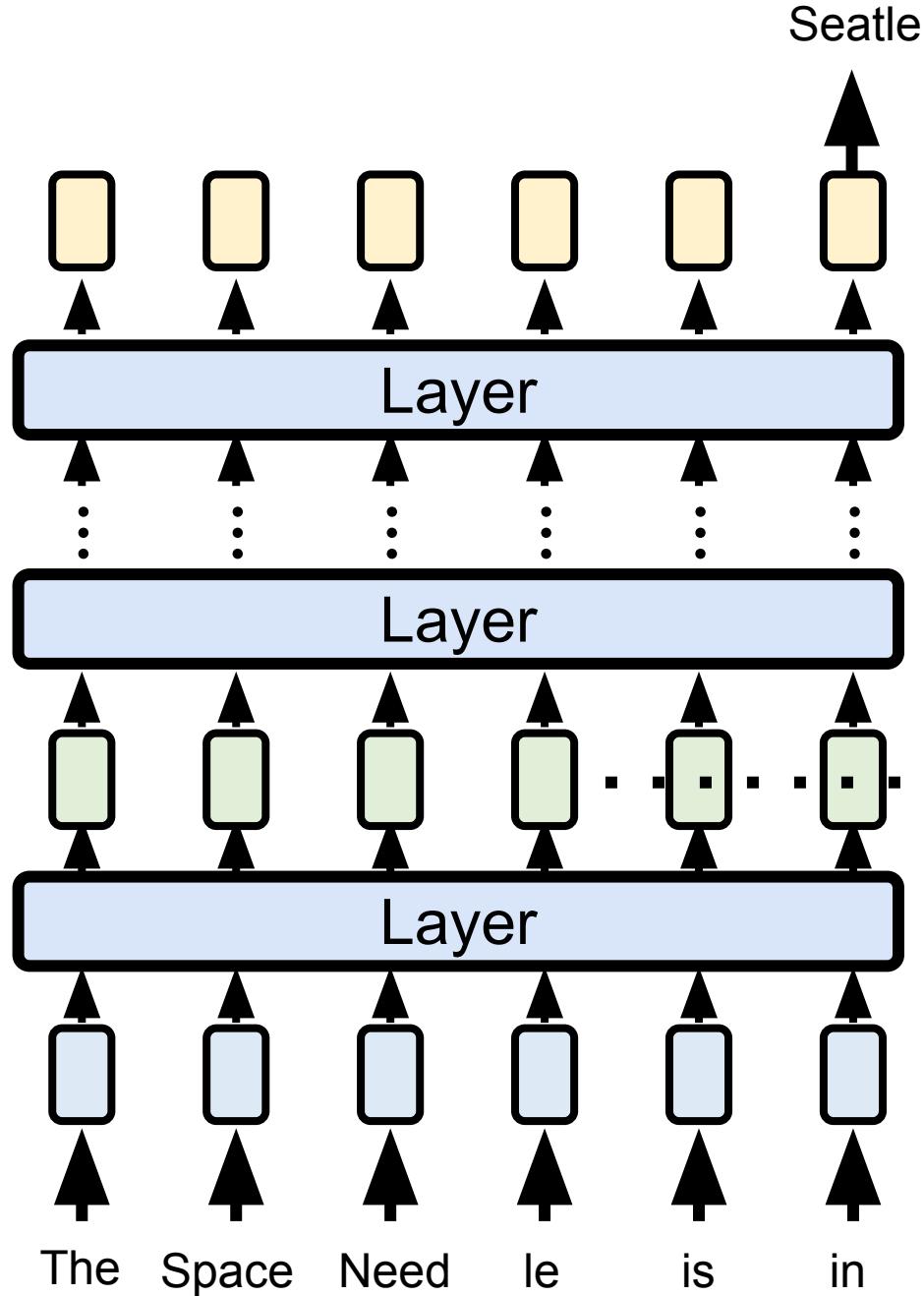
Step 1: 找出類神經網路中跟要
編輯的知識最相關的部分
(參見第三講)



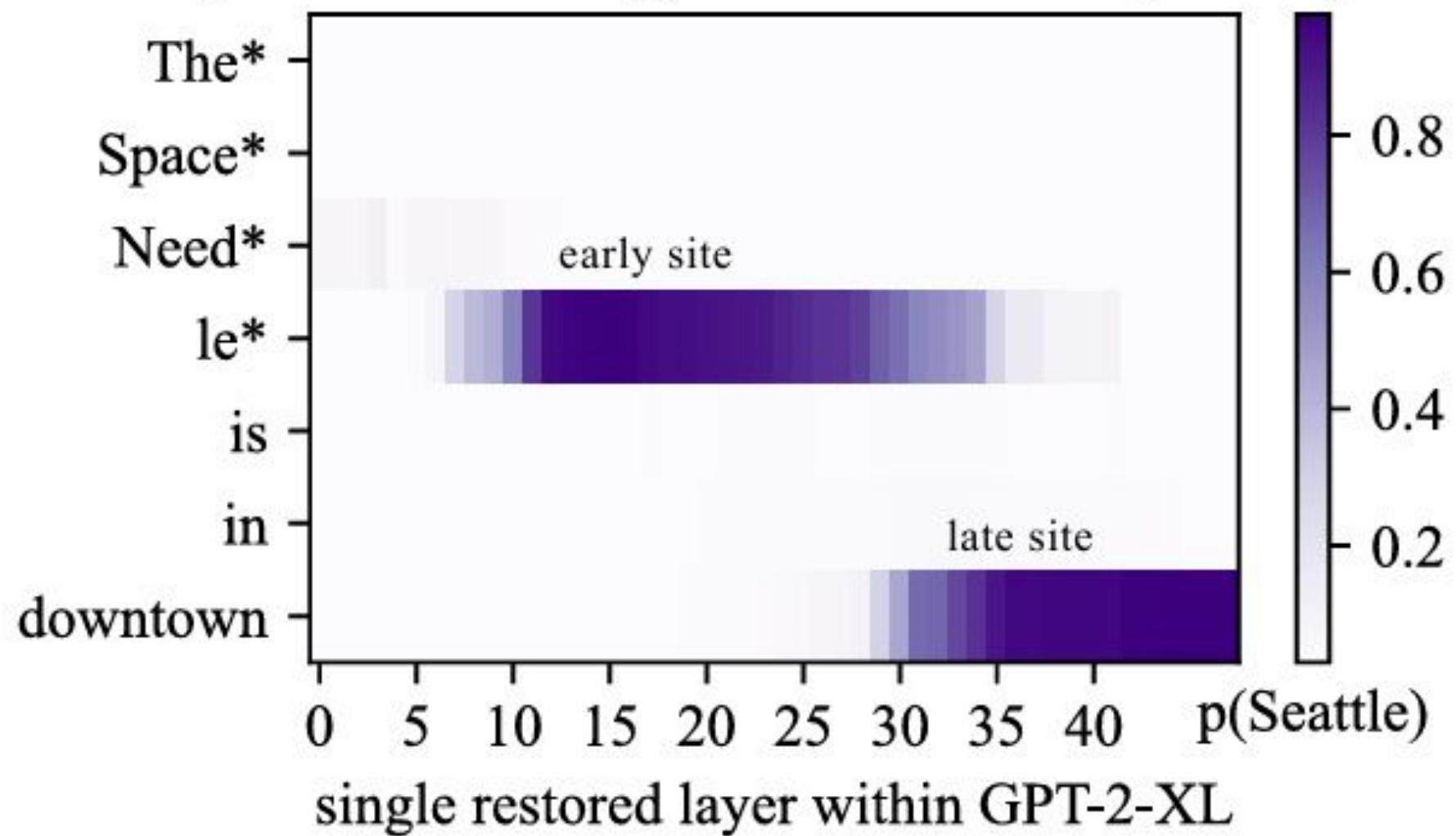
Step 2: 修改該部分的參數
(這就是思想鋼印的原
理)

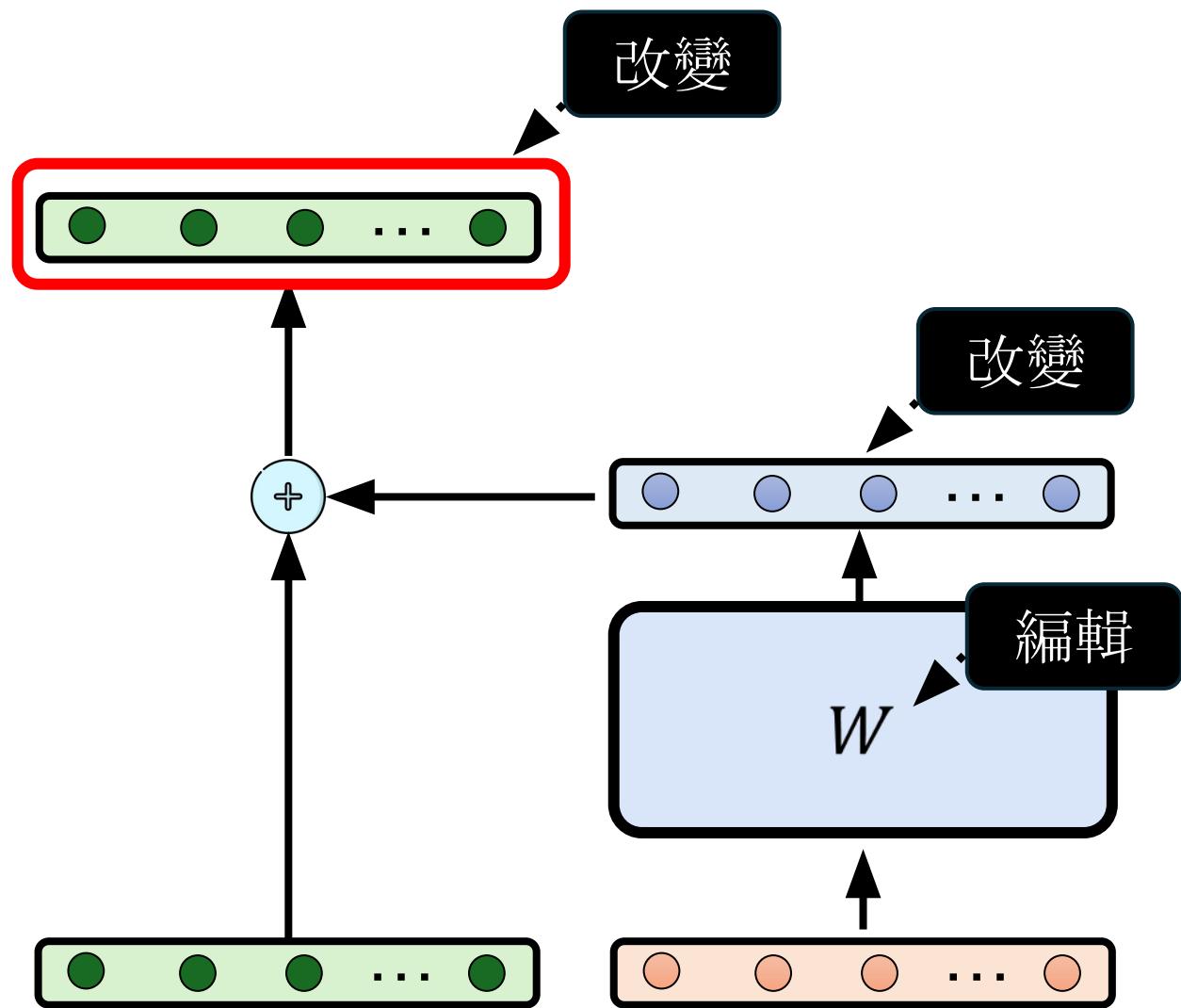
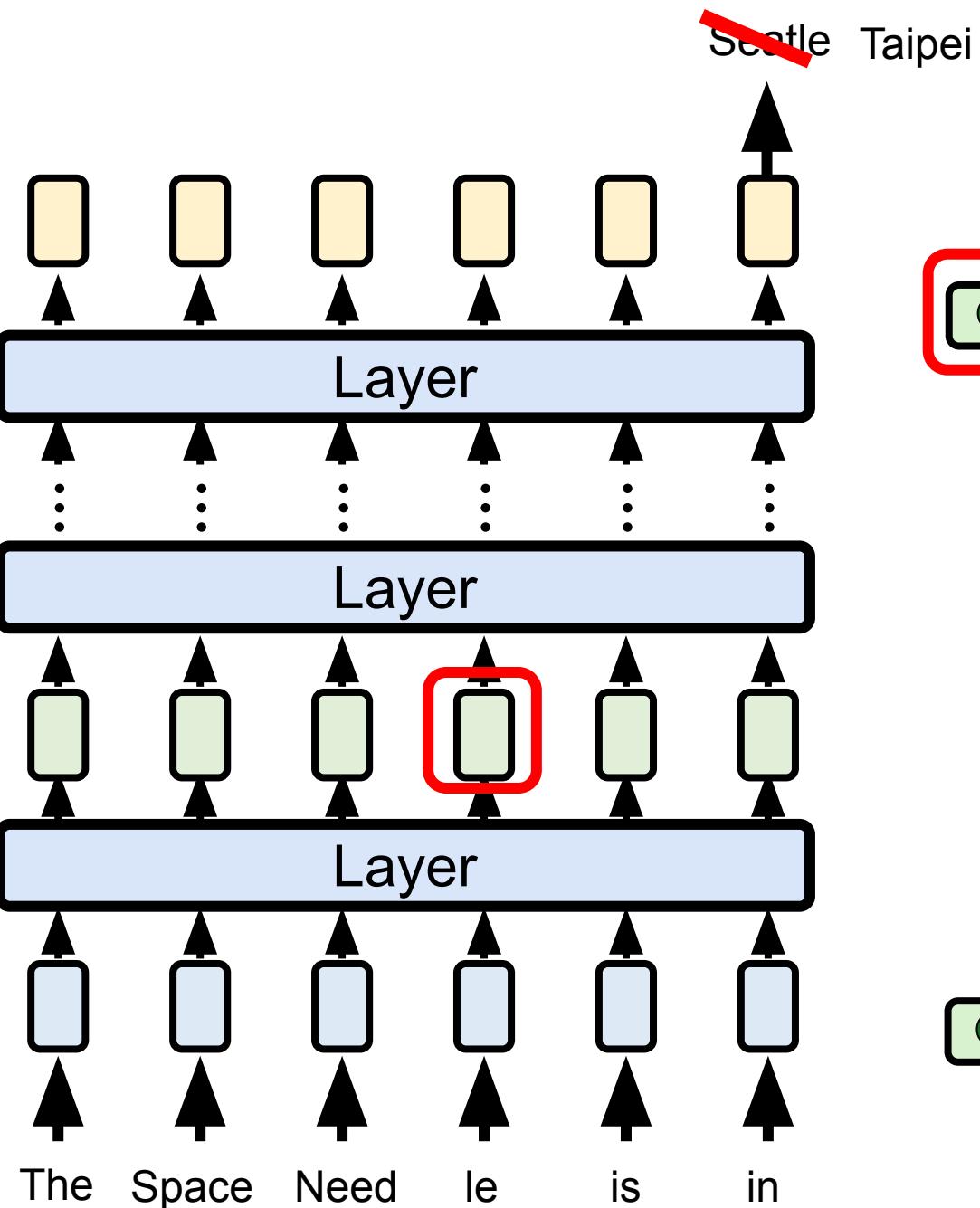


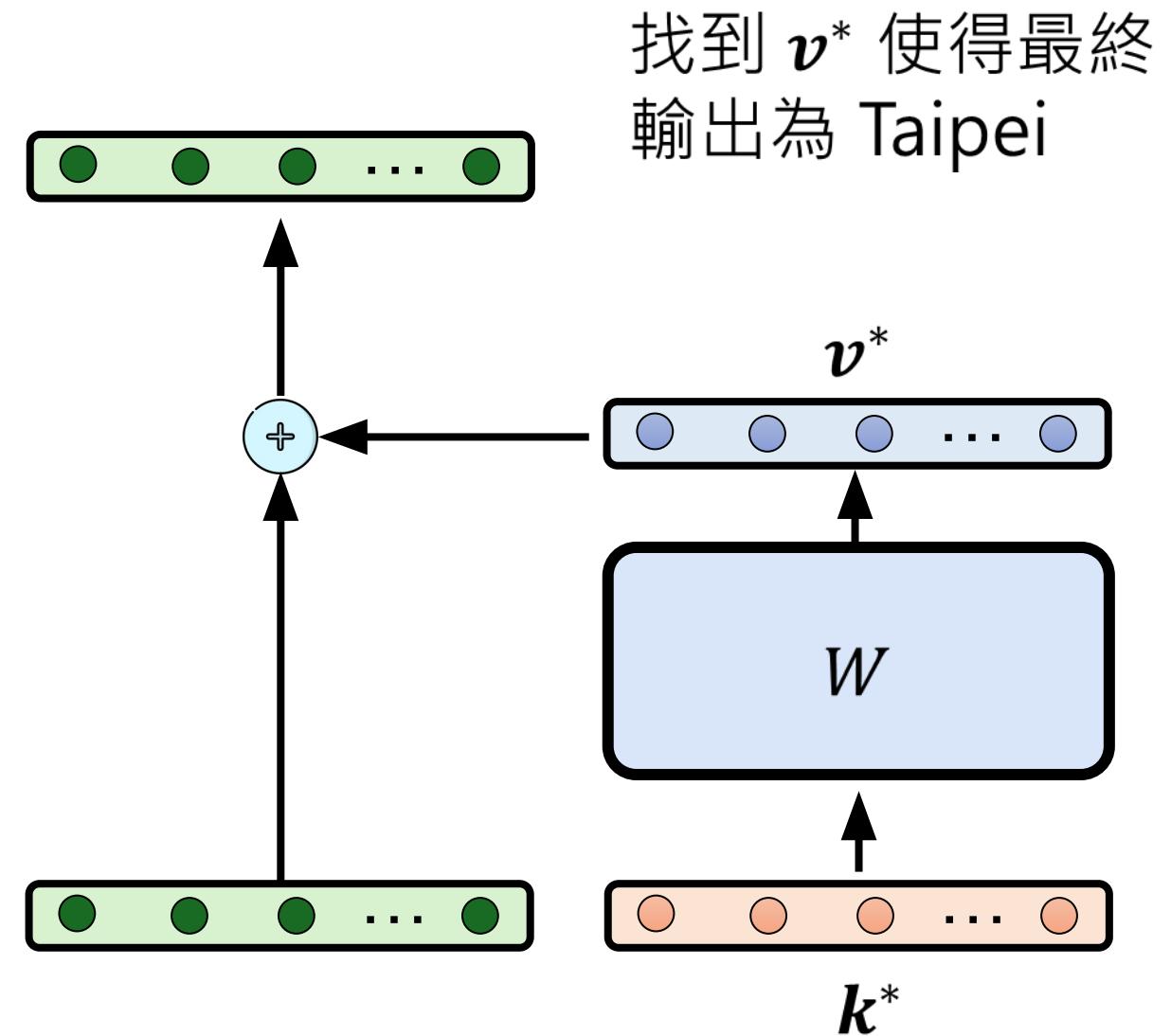
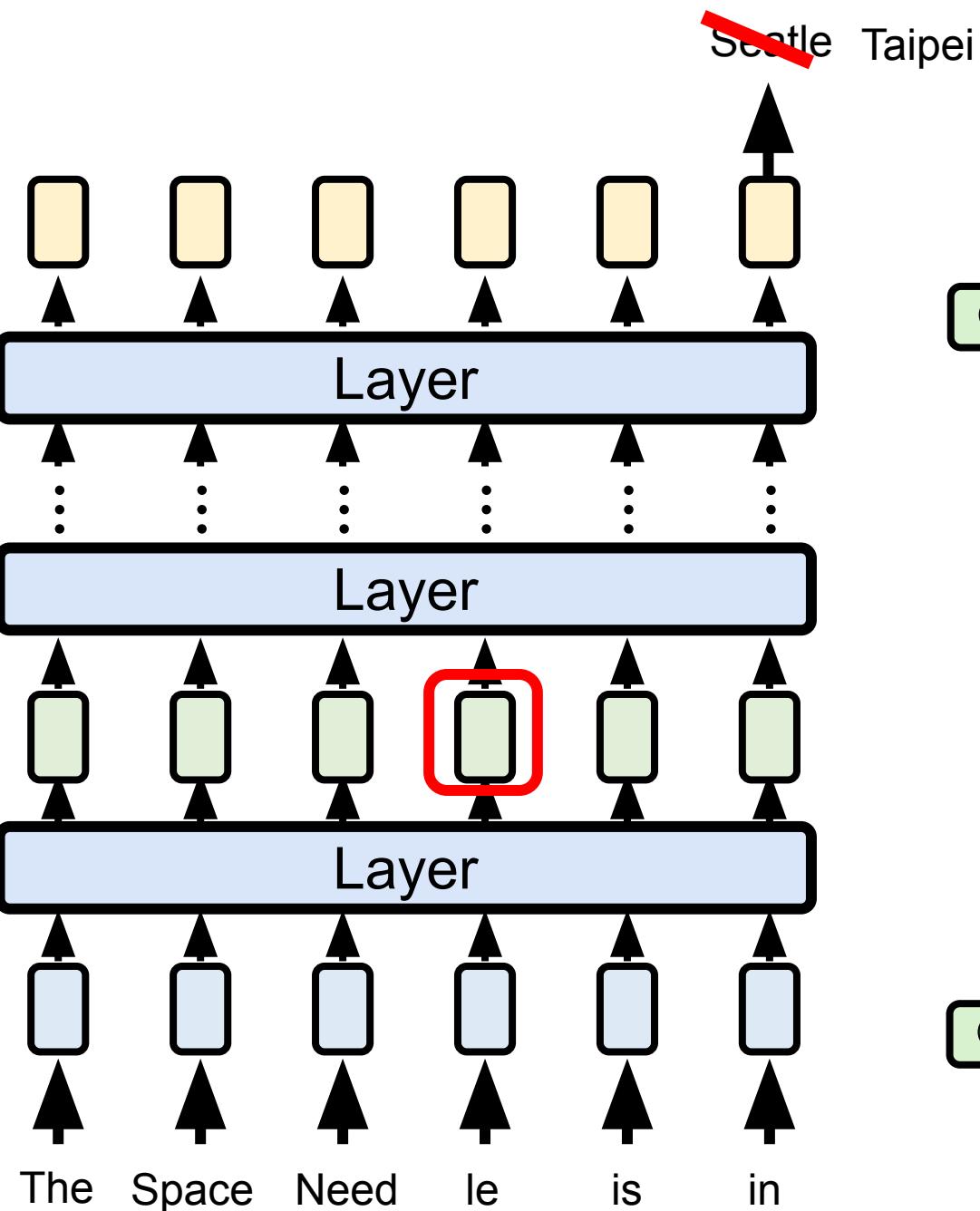


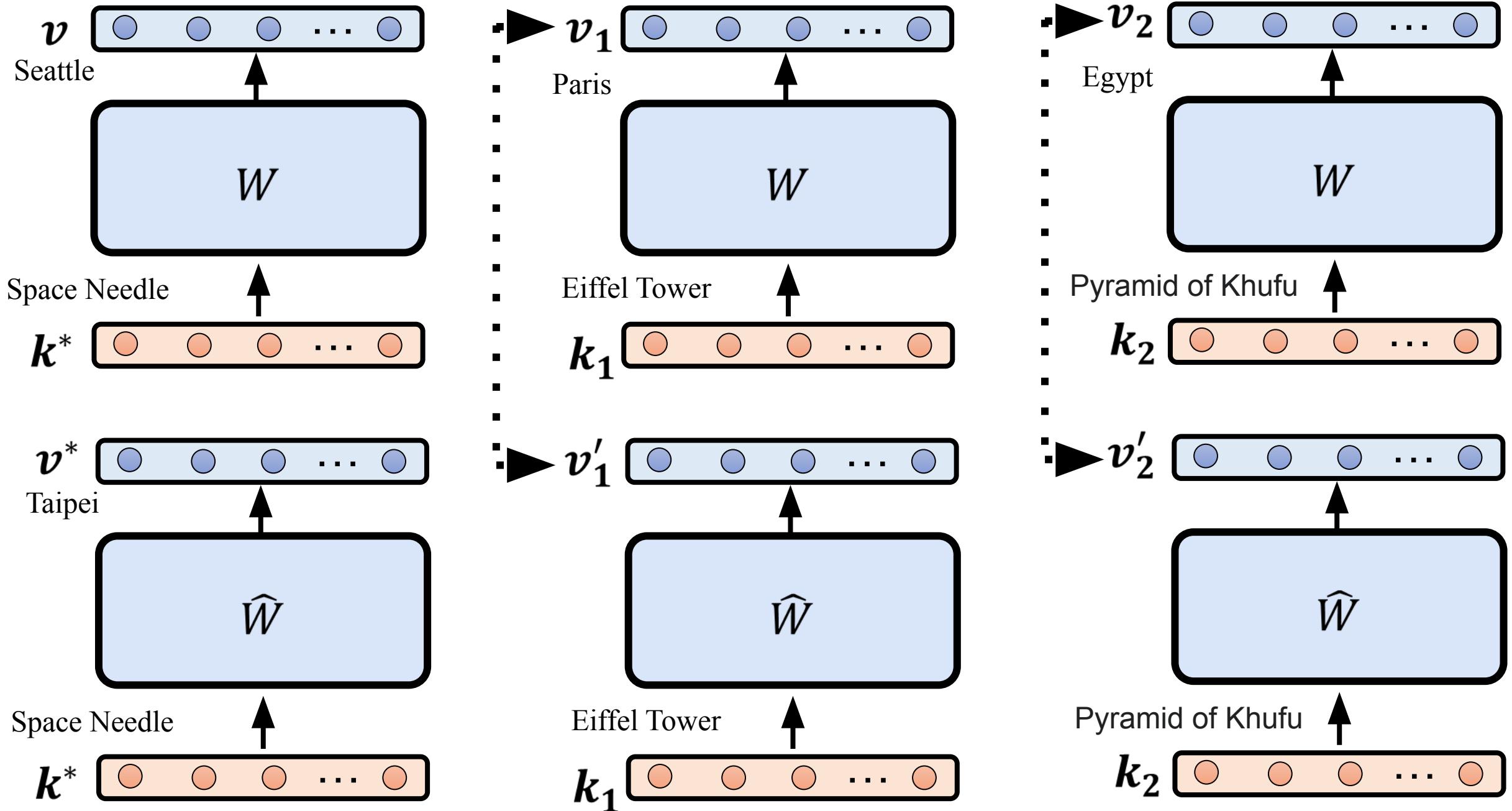


Impact of restoring state after corrupted input





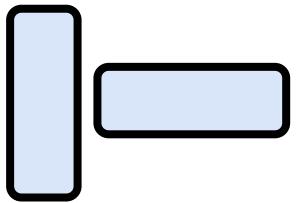




Rank-One Model Editing (ROME)

$$\min_{\hat{W}} \sum_{n=1}^N \|\hat{W}\mathbf{k}_n - \mathbf{v}_n\| \quad \text{such that } \hat{W}\mathbf{k}^* = \mathbf{v}^*$$

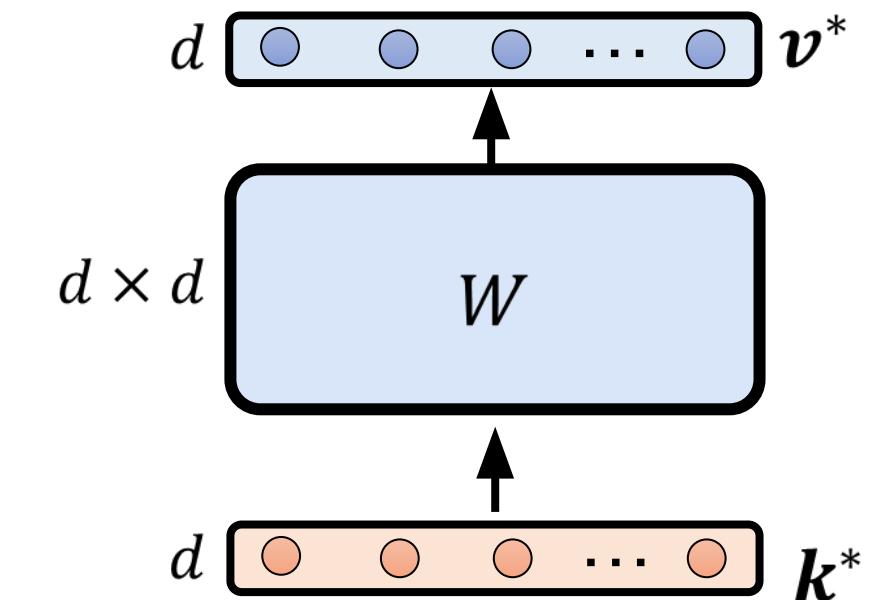
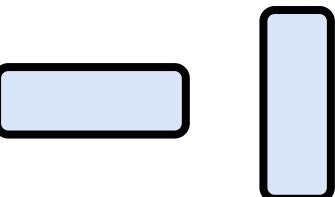
$$\hat{W} = W + \Lambda(C^{-1}\mathbf{k}^*)^T$$



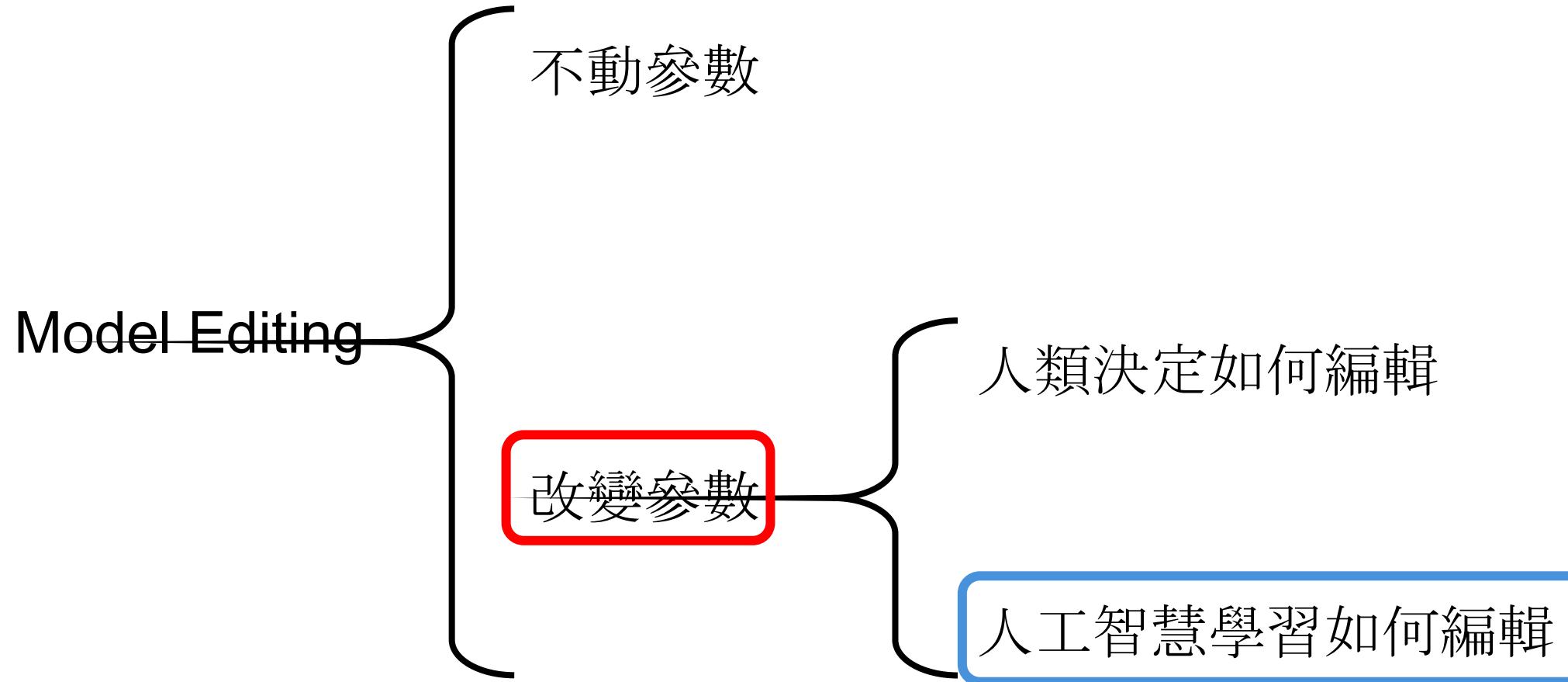
$$C = KK^T \quad d \times d$$

$$K = [\mathbf{k}_1 \quad \mathbf{k}_2 \quad \dots \quad \mathbf{k}_n] \quad d \times n$$

$$\Lambda = \frac{1}{\lambda} (\mathbf{v}^* - W\mathbf{k}^*) \quad \lambda = (C^{-1}\mathbf{k}^*)^T \mathbf{k}^*$$

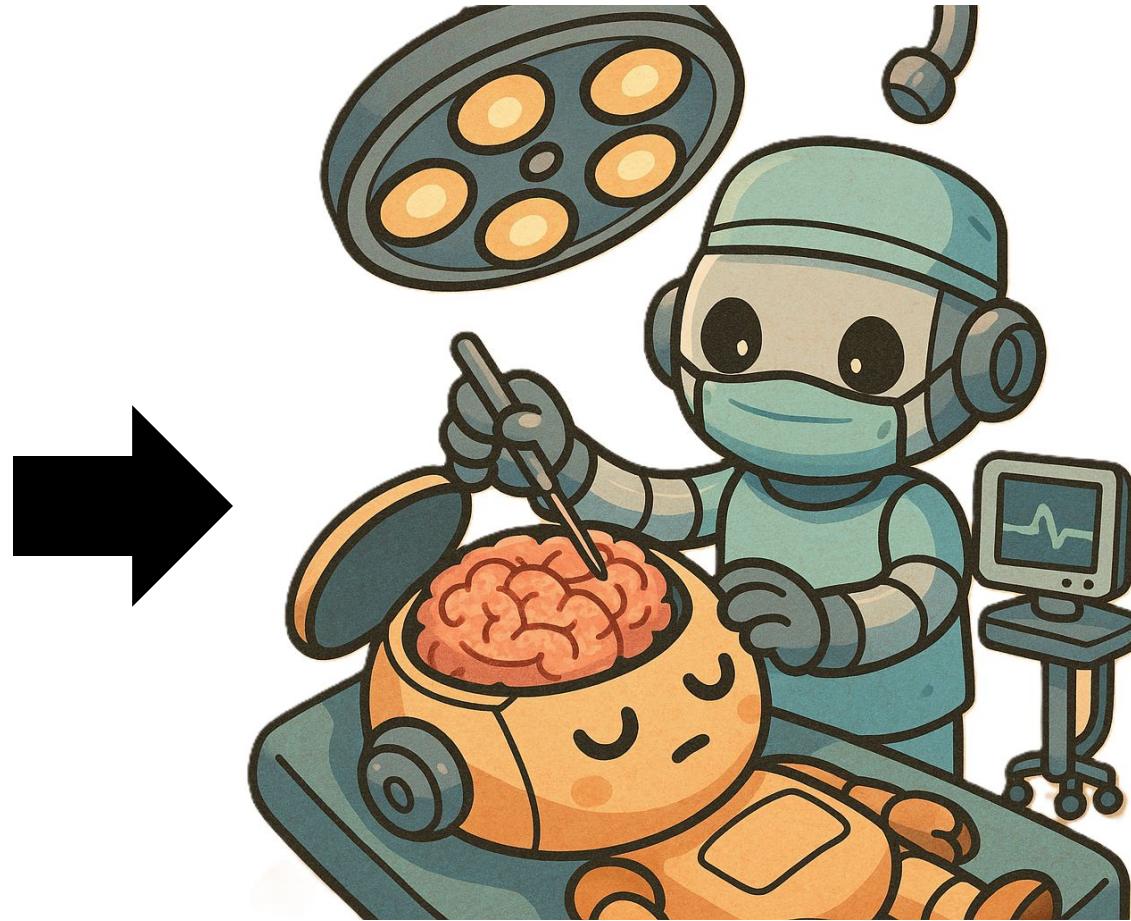


Model Editing 常見方法

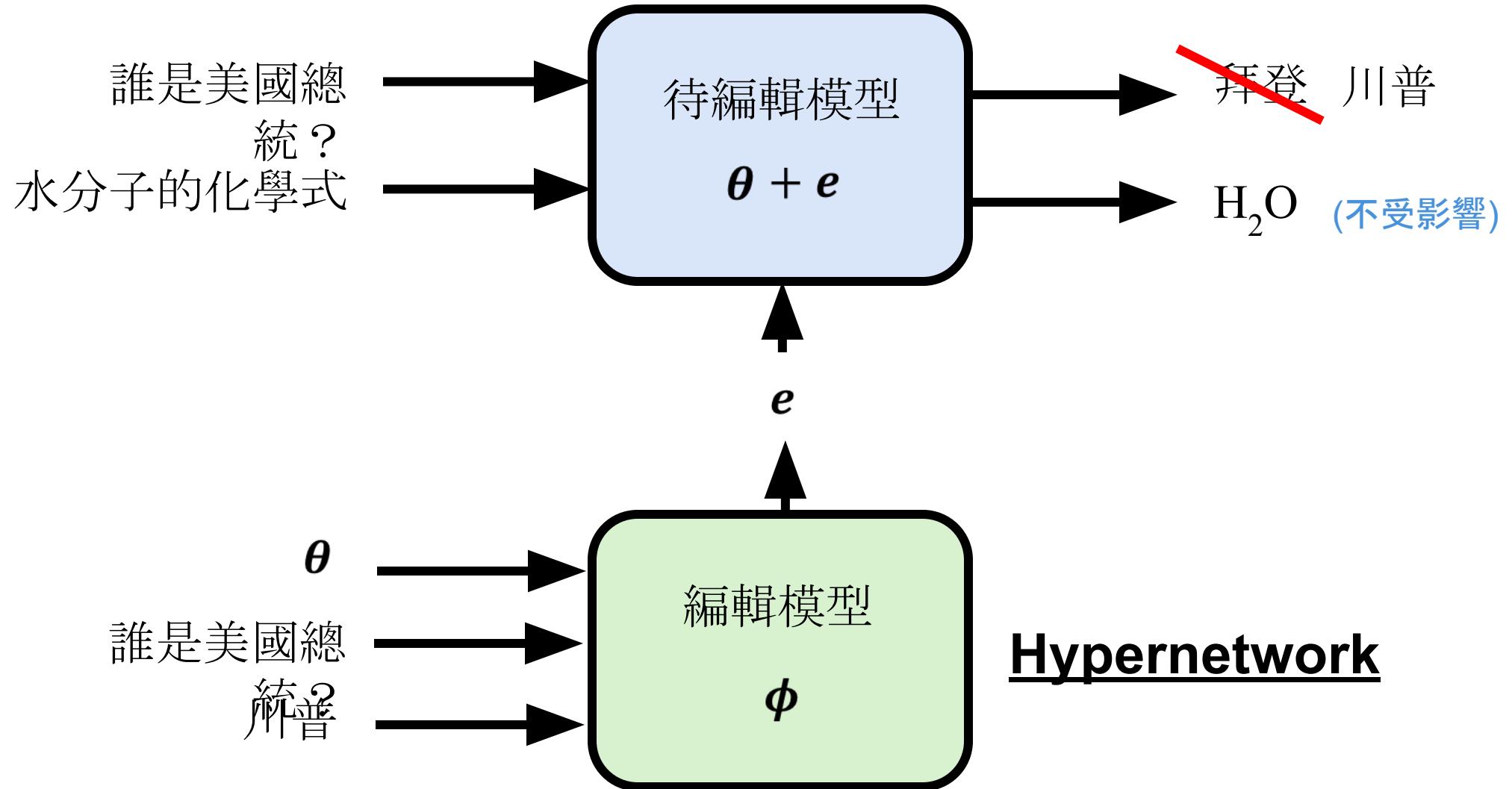




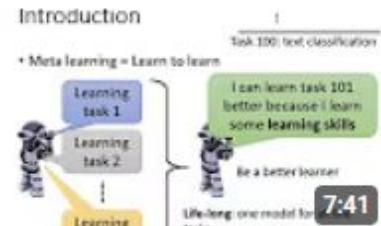
由人類決定要如何
進行編輯



讓另外一個人工智慧
學習如何編輯

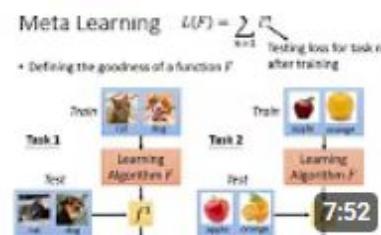


Meta Learning



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

More about Meta Learning
Hung-yi Lee
10:22

Meta Learning 完整介紹請見《機器學習2019》

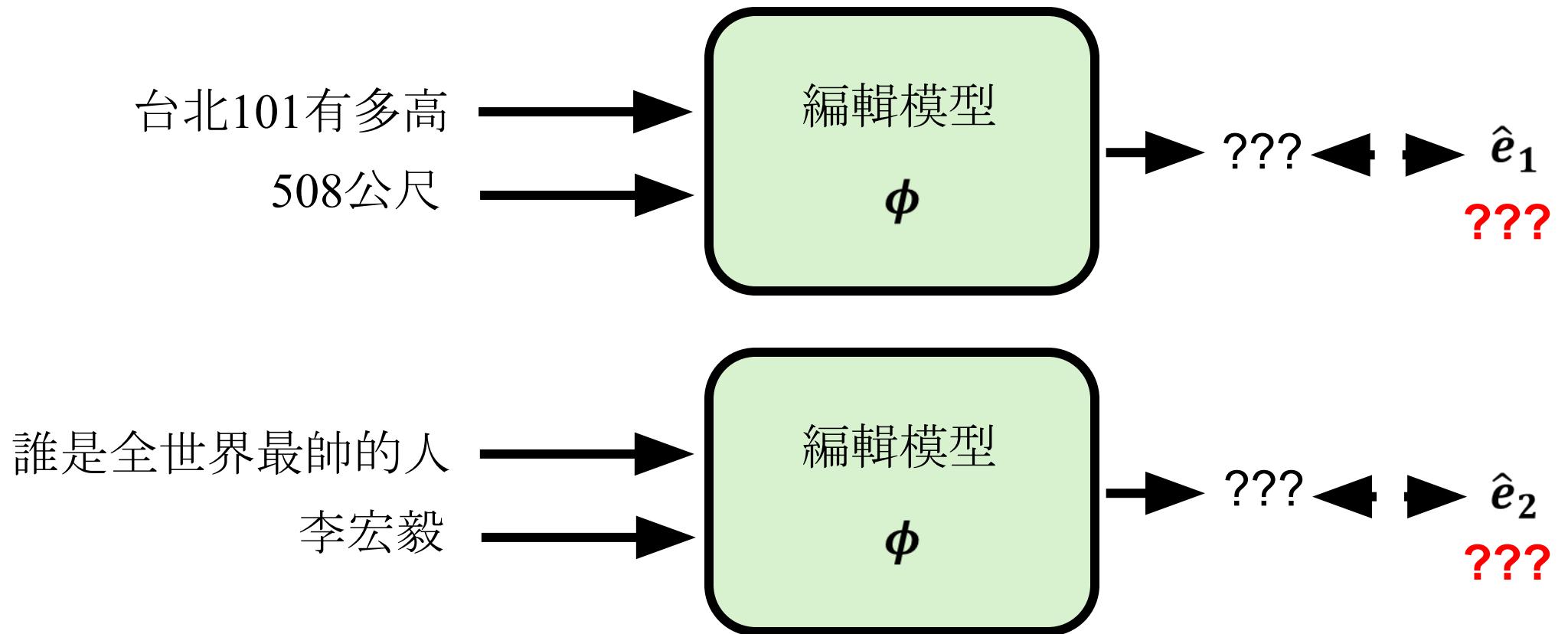
https://www.youtube.com/playlist?list=PLJV_el3uVTsOK_ZK5L0lv_EQoL1JefRL4

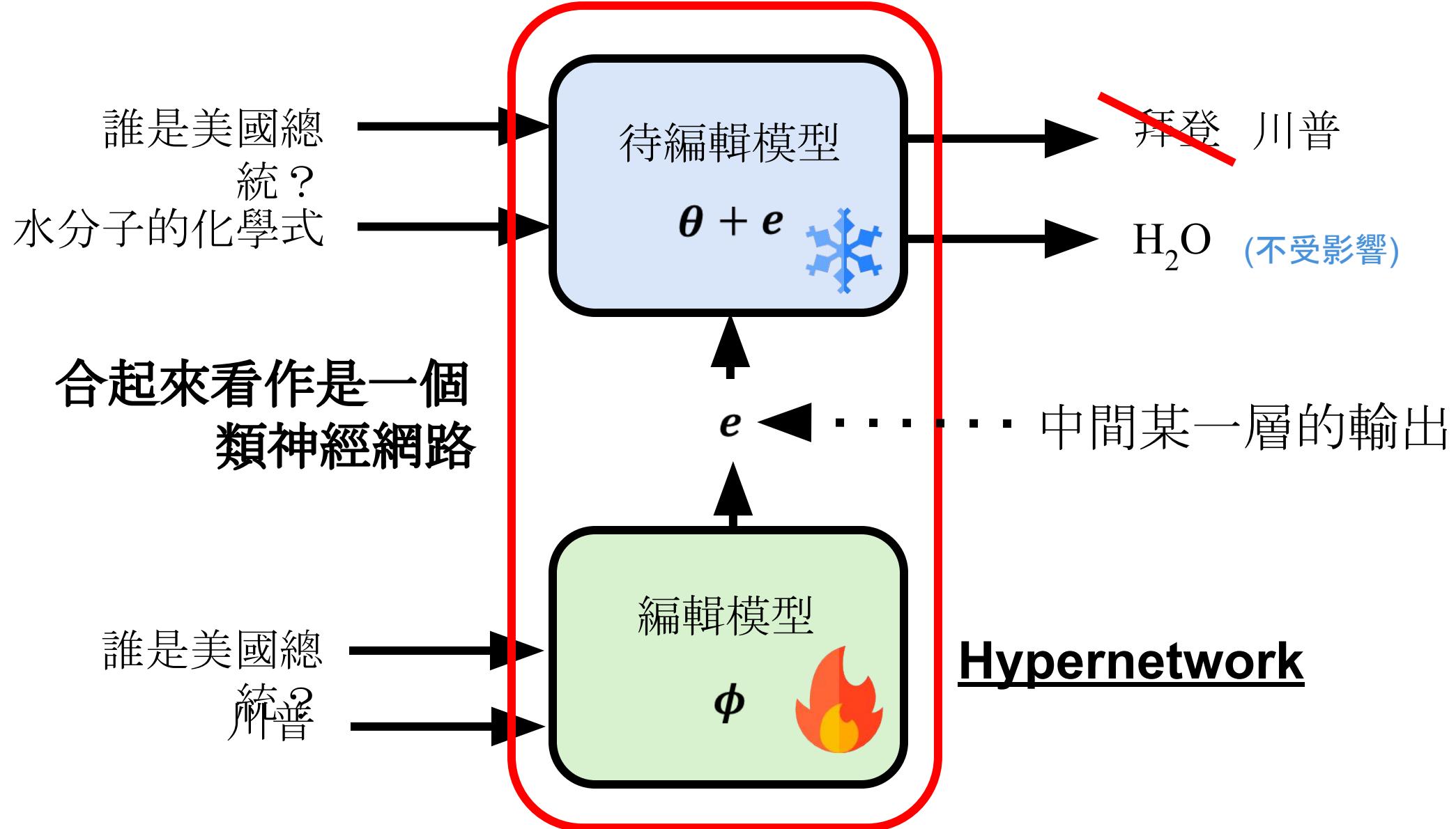
【機器學習 2022】各種奇葩的元學習 (Meta Learning) 用法

https://youtu.be/QNfymMRUg3M?si=GQP2H_pGyqLR6cWI

如何訓練 Hypernetwork ?

待編輯模型
 $\theta + e$

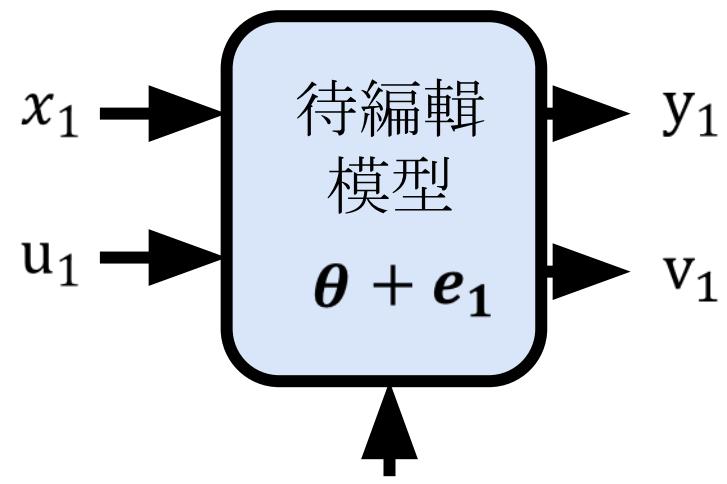




Training

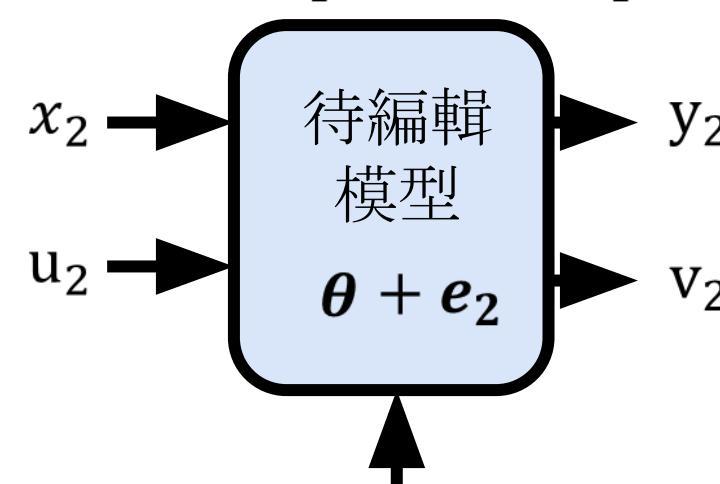
輸入 x_1 、改成輸出 y_1

輸入 u_1 、維持輸出 v_1



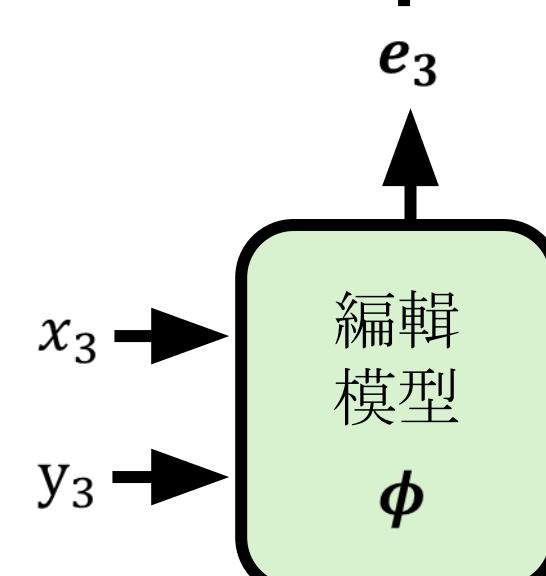
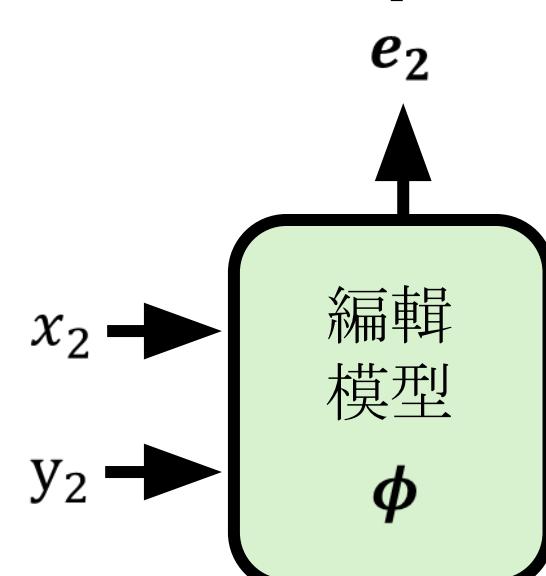
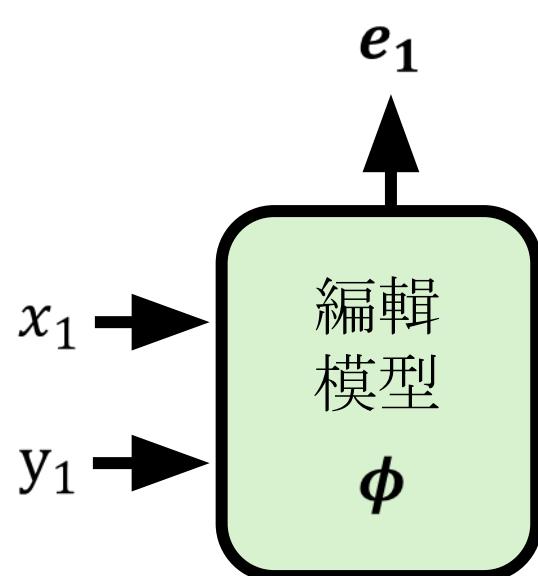
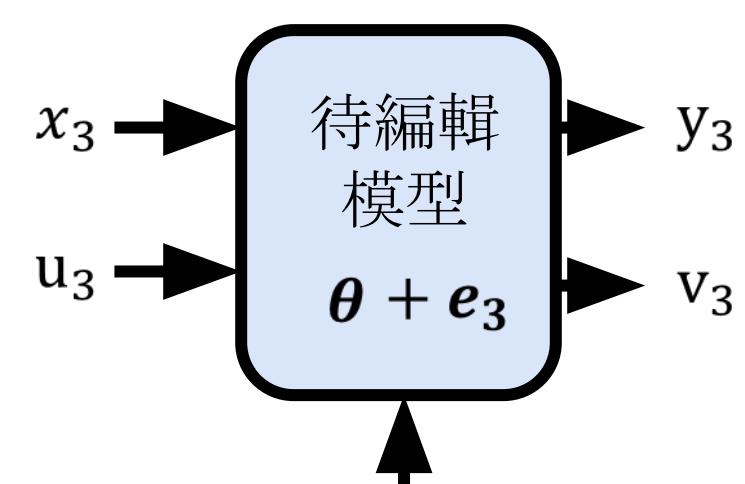
輸入 x_2 、改成輸出 y_2

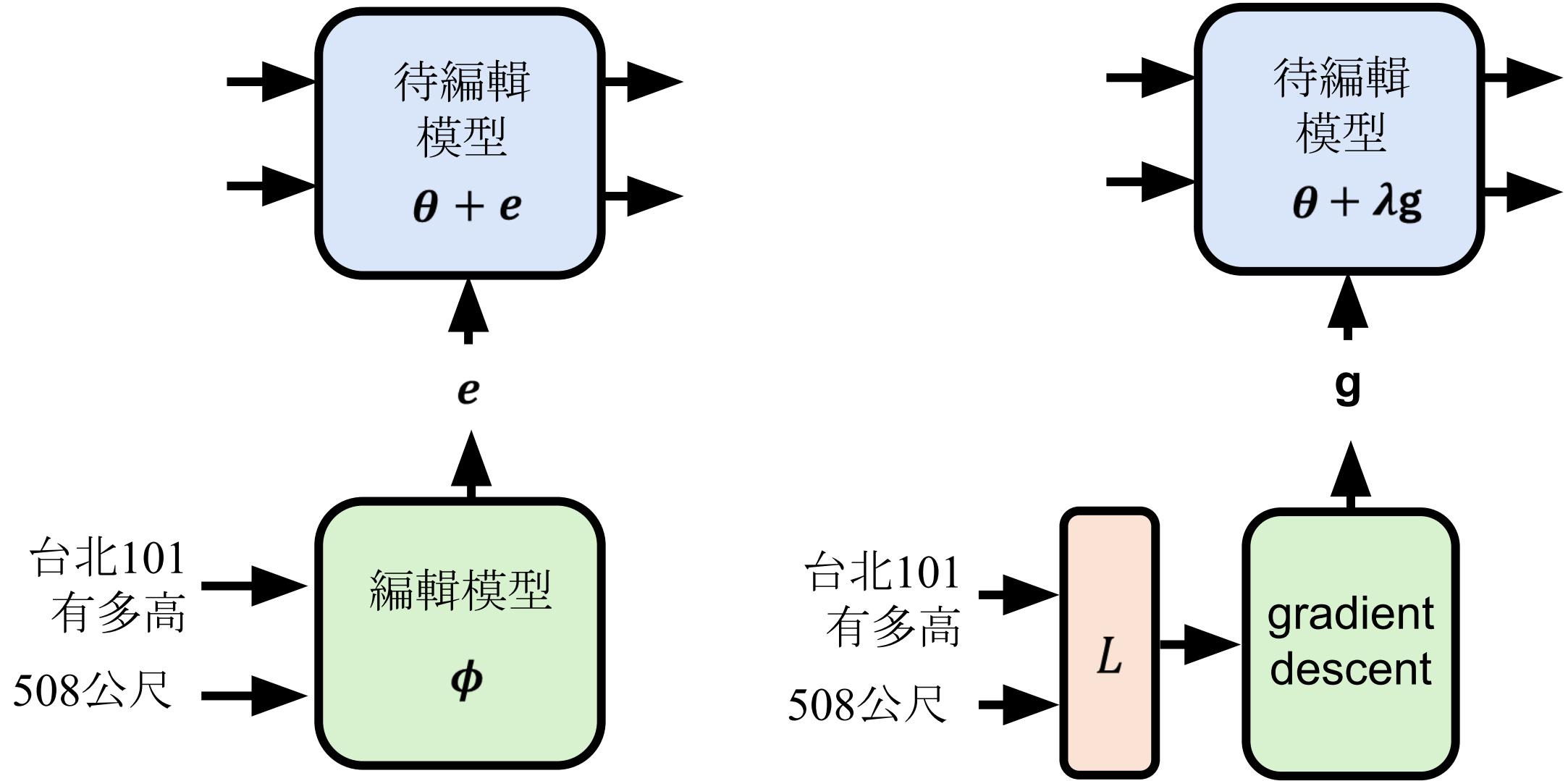
輸入 u_2 、維持輸出 v_2



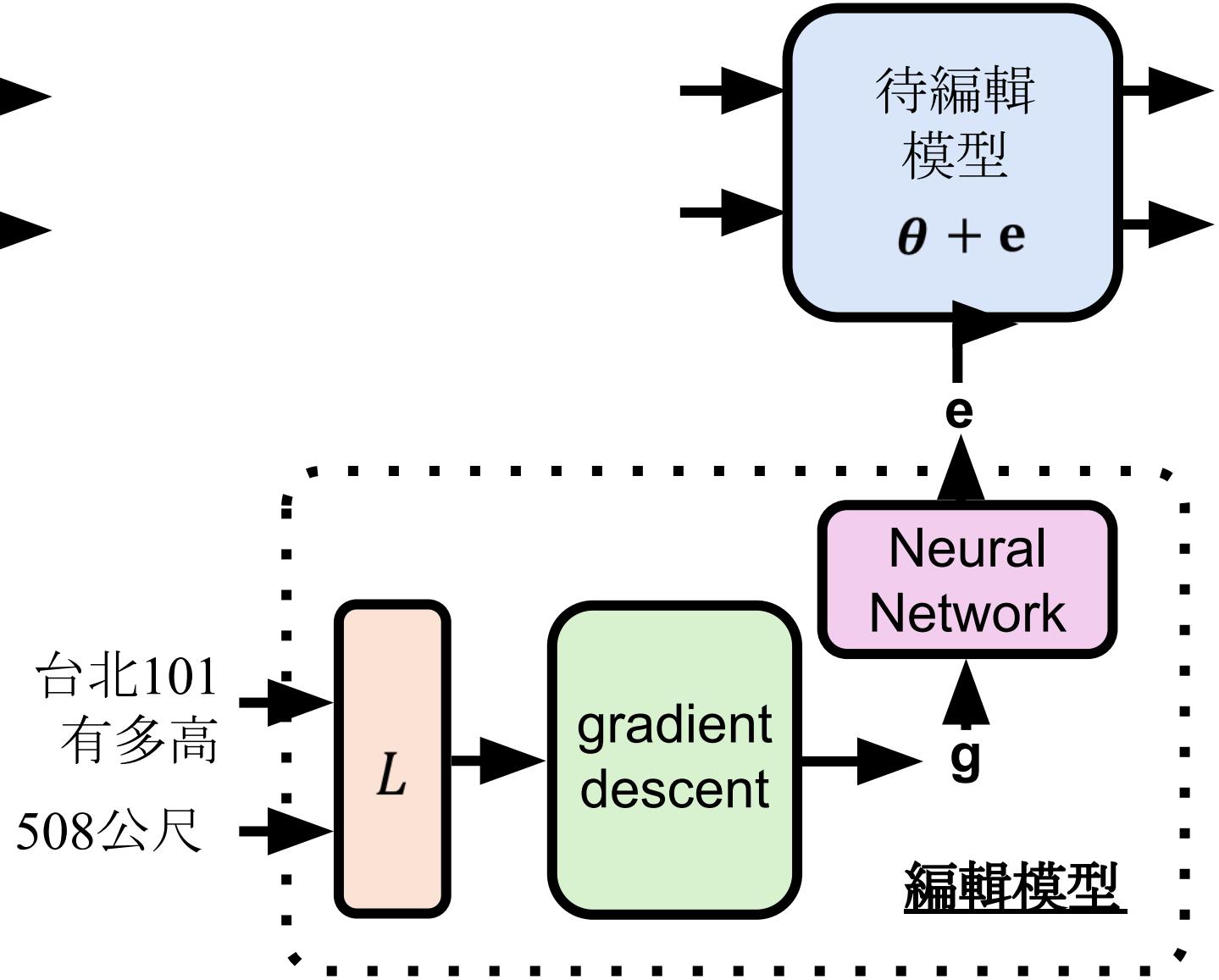
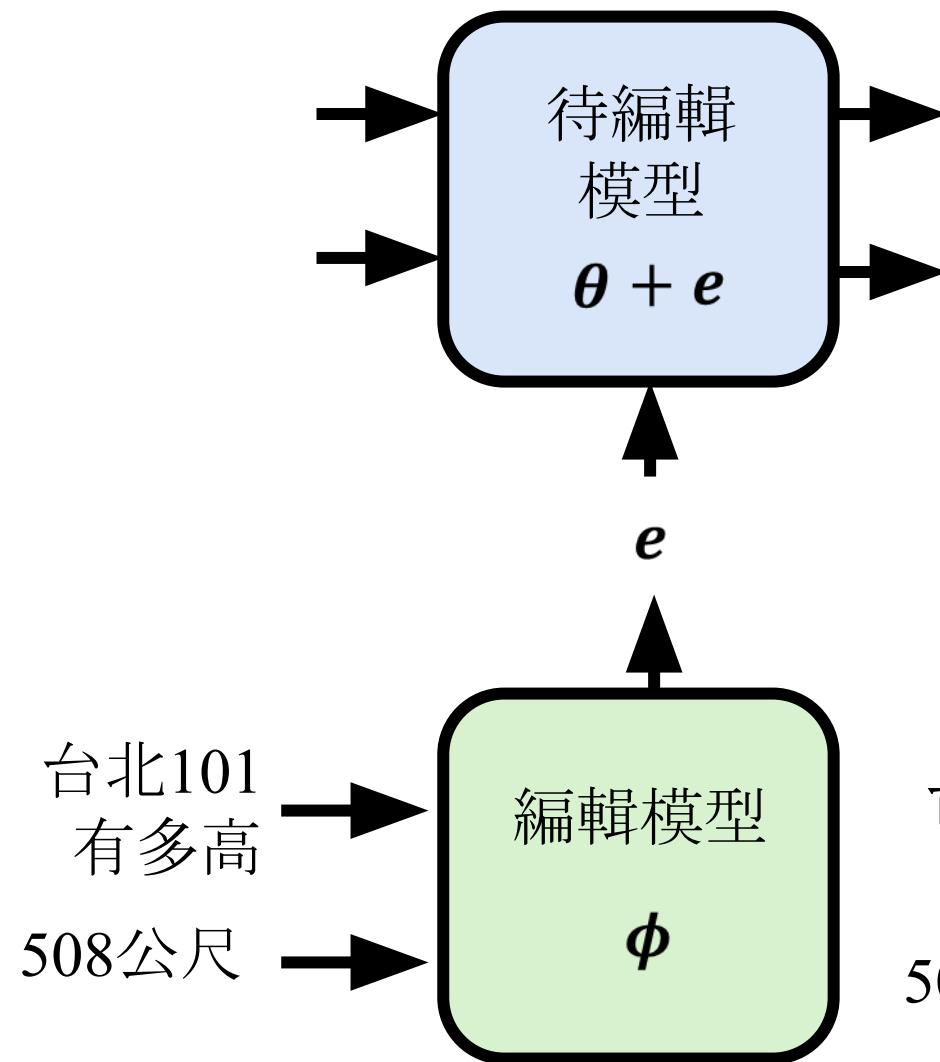
Testing

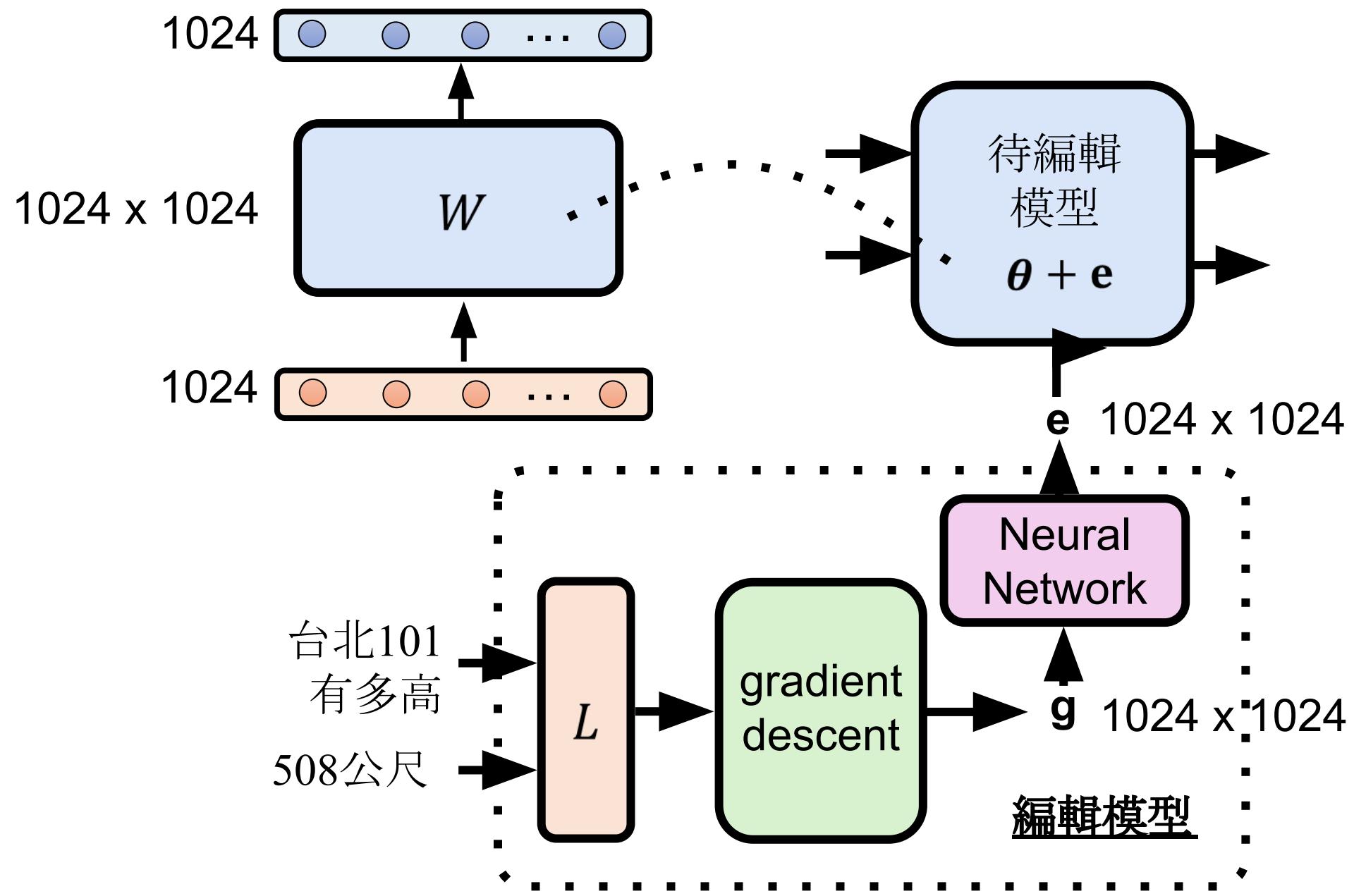
輸入 x_3 、改成輸出 y_3





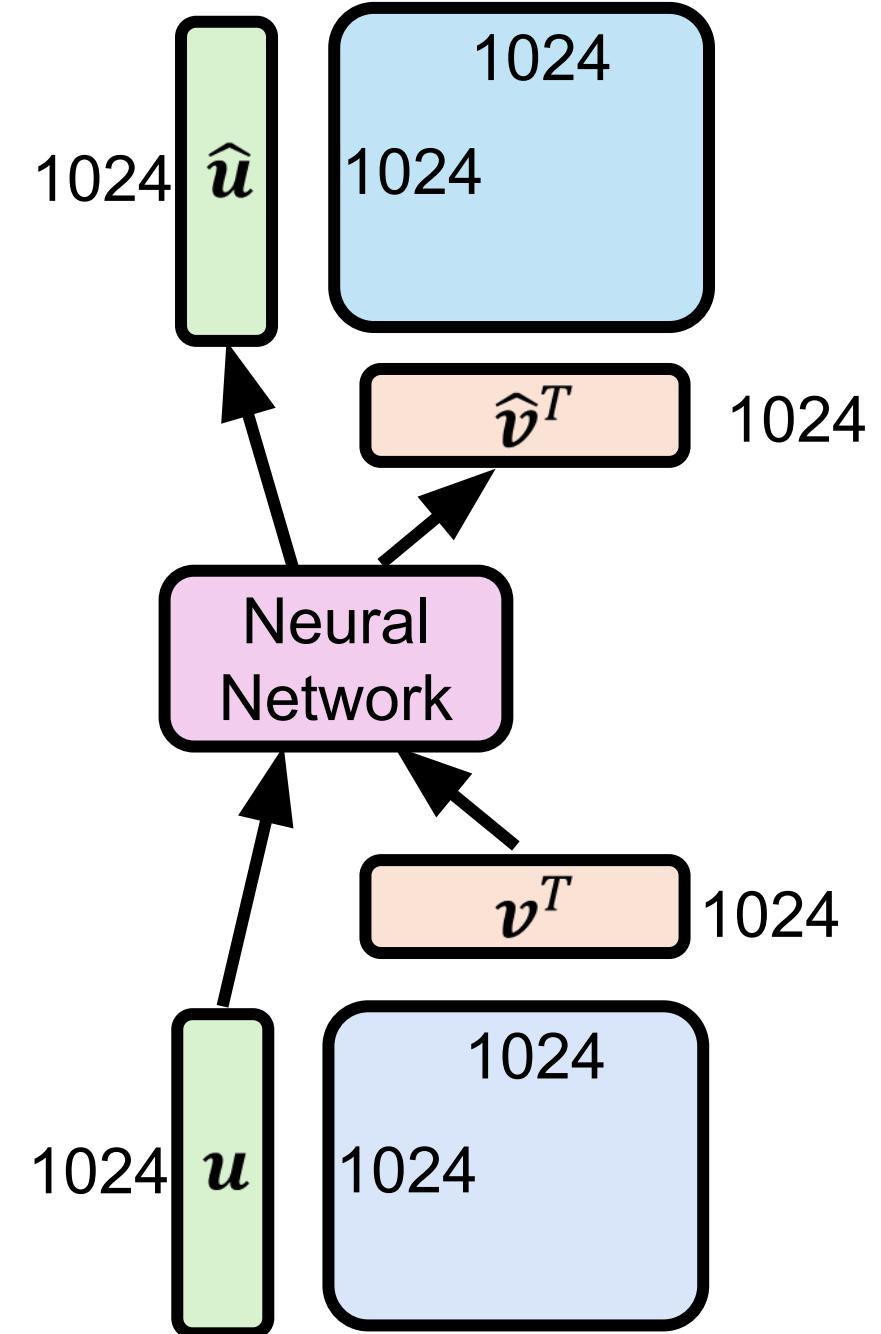
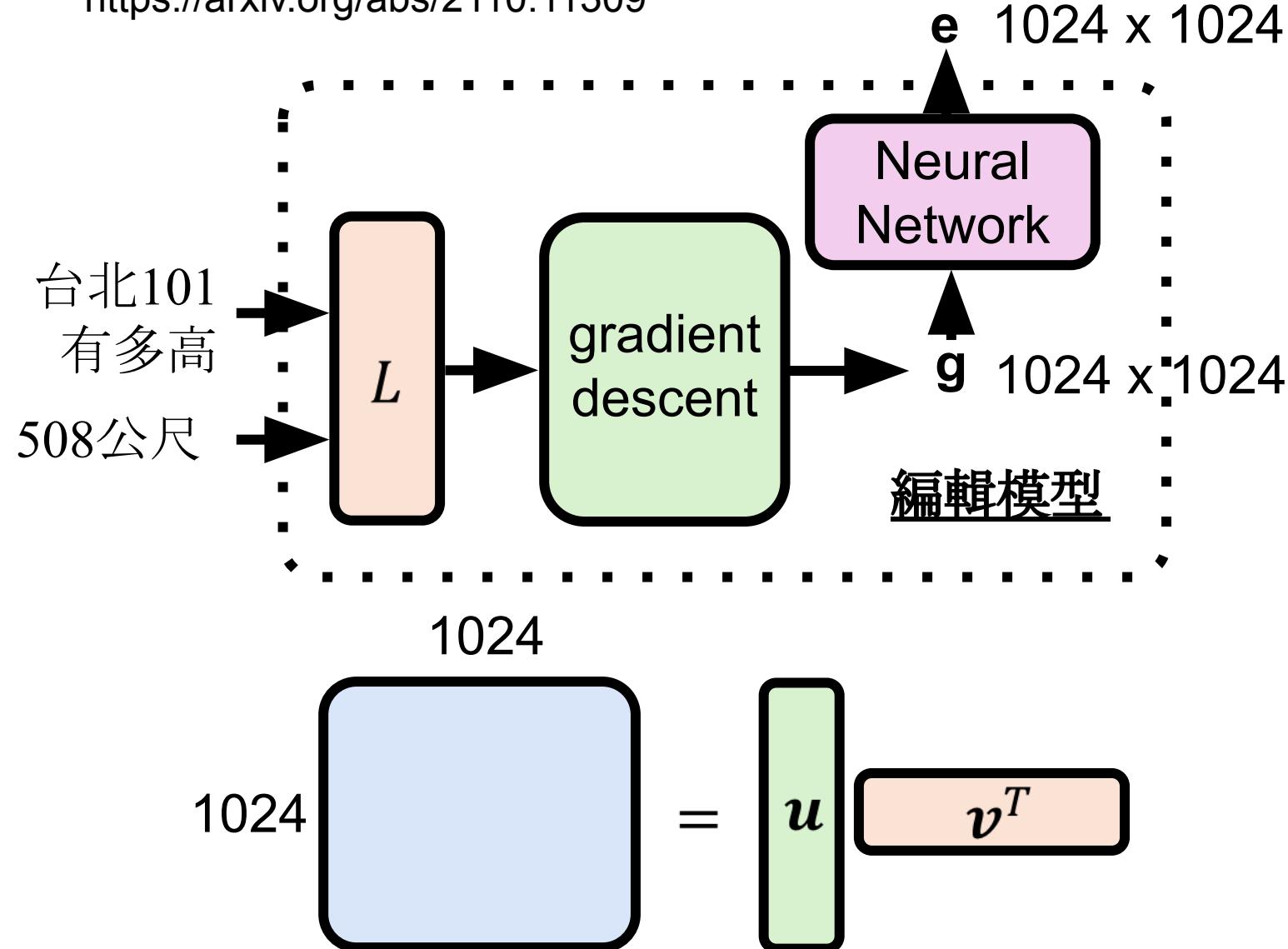
<https://arxiv.org/abs/2104.08164>
<https://arxiv.org/abs/2110.11309>



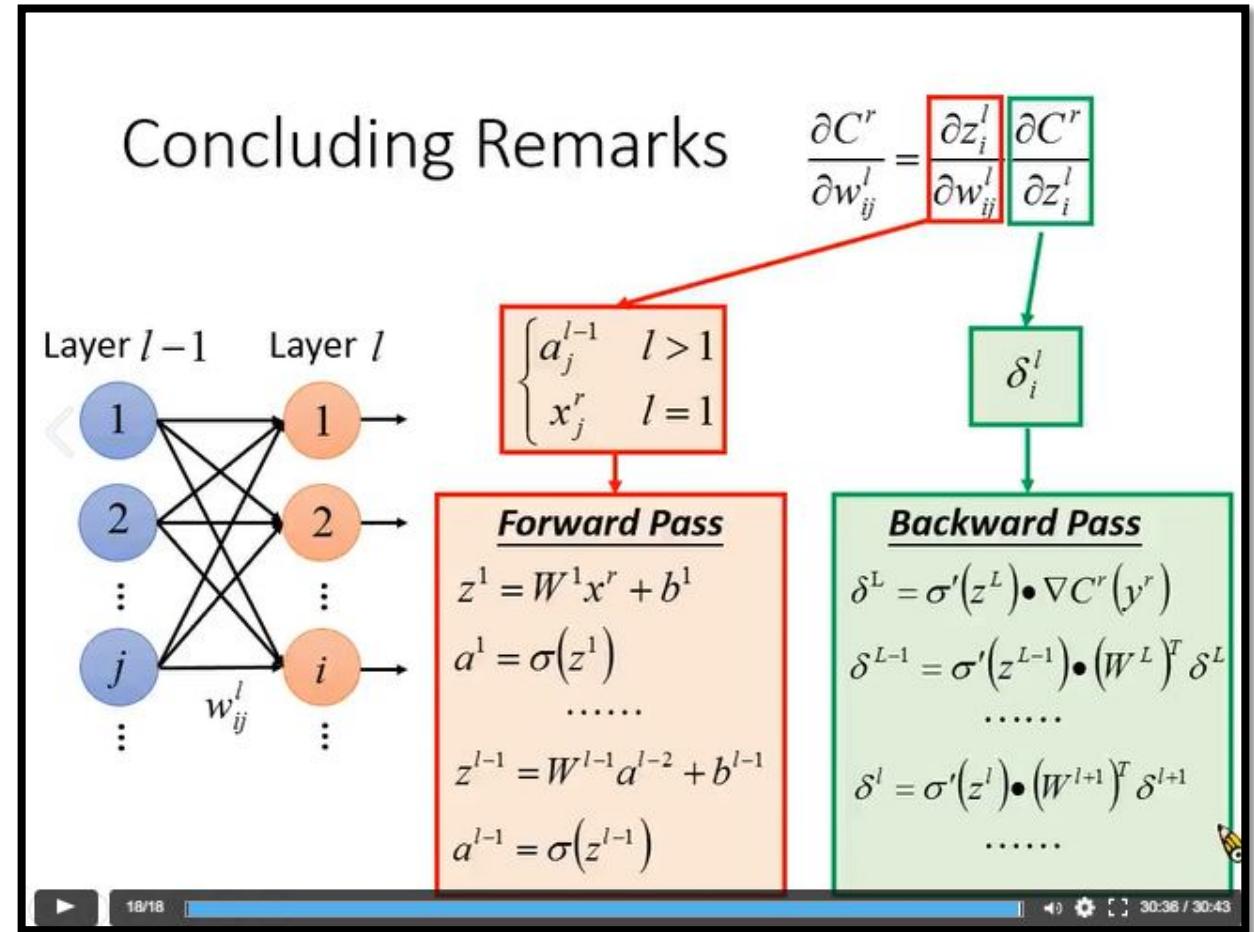


MEND

<https://arxiv.org/abs/2110.11309>

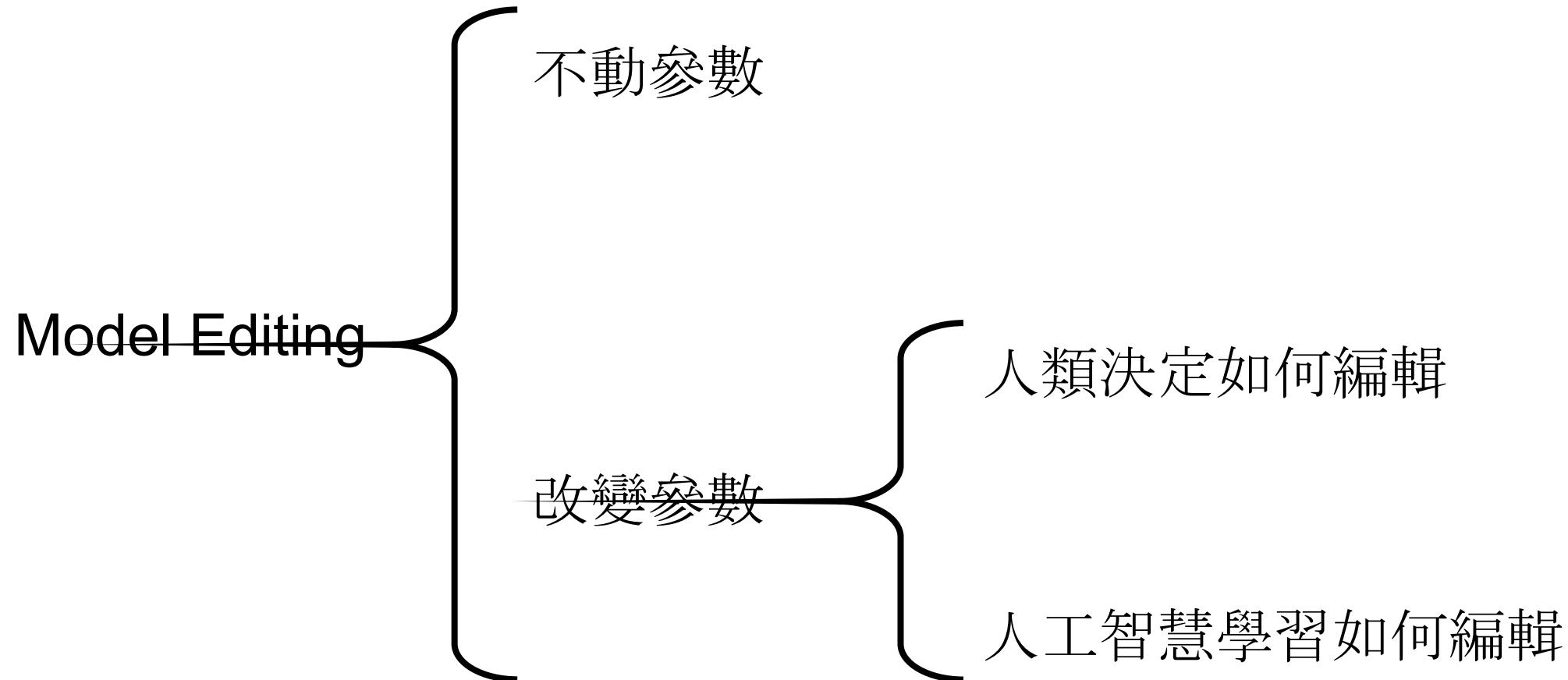


$$1024 \times 1024 = u v^T$$



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 ...



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 Wang¹, Zekun Xi¹, Shengyu Mao¹, Jintian Zhang¹, Yuansheng Ni¹, Siyuan Cheng¹,
Ziwen Xu¹, Xin Xu¹, Jia-Chen Gu¹, Yong Jiang¹, Pengjun Xie¹, Fei Huang¹, Lei Liang¹, Zhiqiang Zhang¹,
Xiaowei Zhu¹, Jun Zhou¹, Huajun Chent¹

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

⁵Alibaba Group

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