自然語言處理與應用
Natural Language Processing and Applications

Mixture of Experts (MoE)

Instructor: 林英嘉 (Ying-Jia Lin)
2025/06/01

Course GitHub

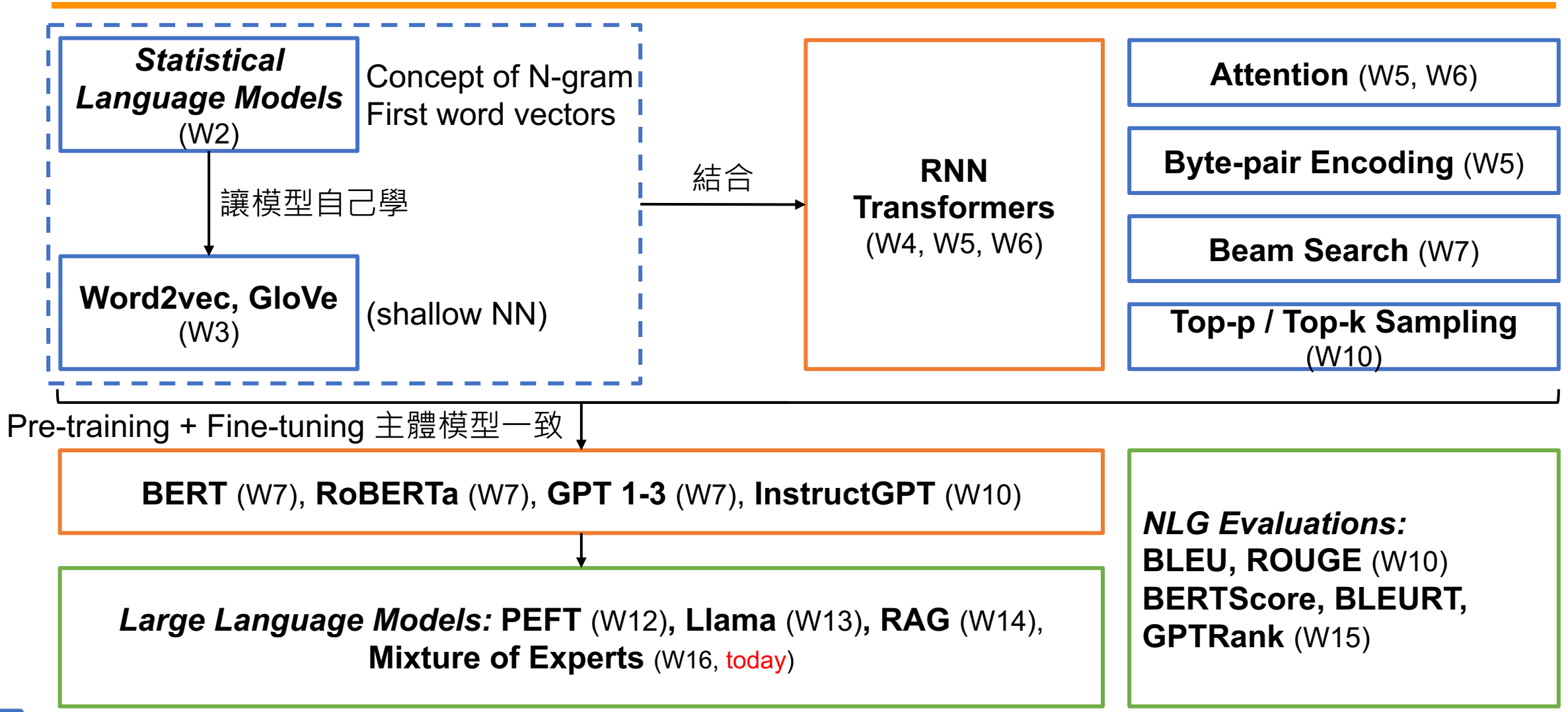
Slido # NLP_0601

Outline

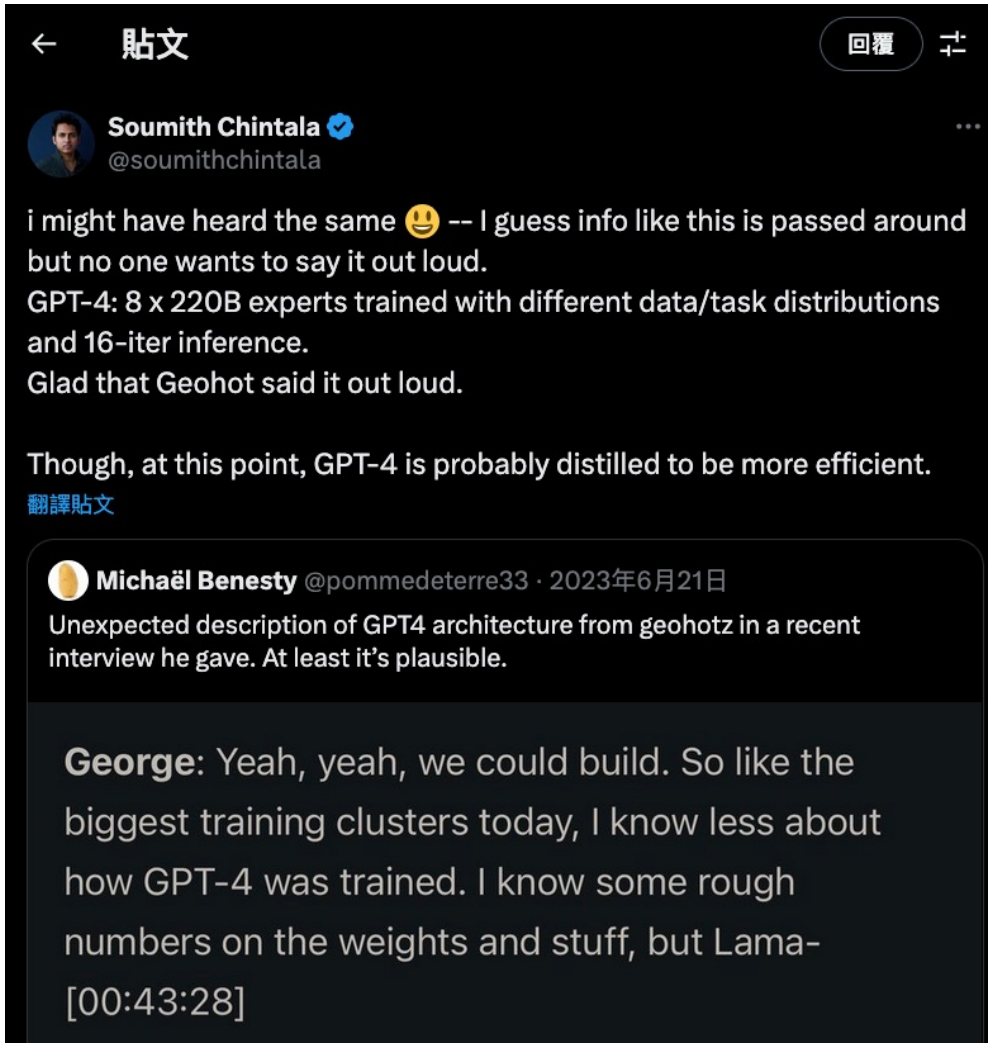
- 學期主題回顧
- Introduction to Mixture of Experts

學期主題回顧

Road Map of Natural Language Processing

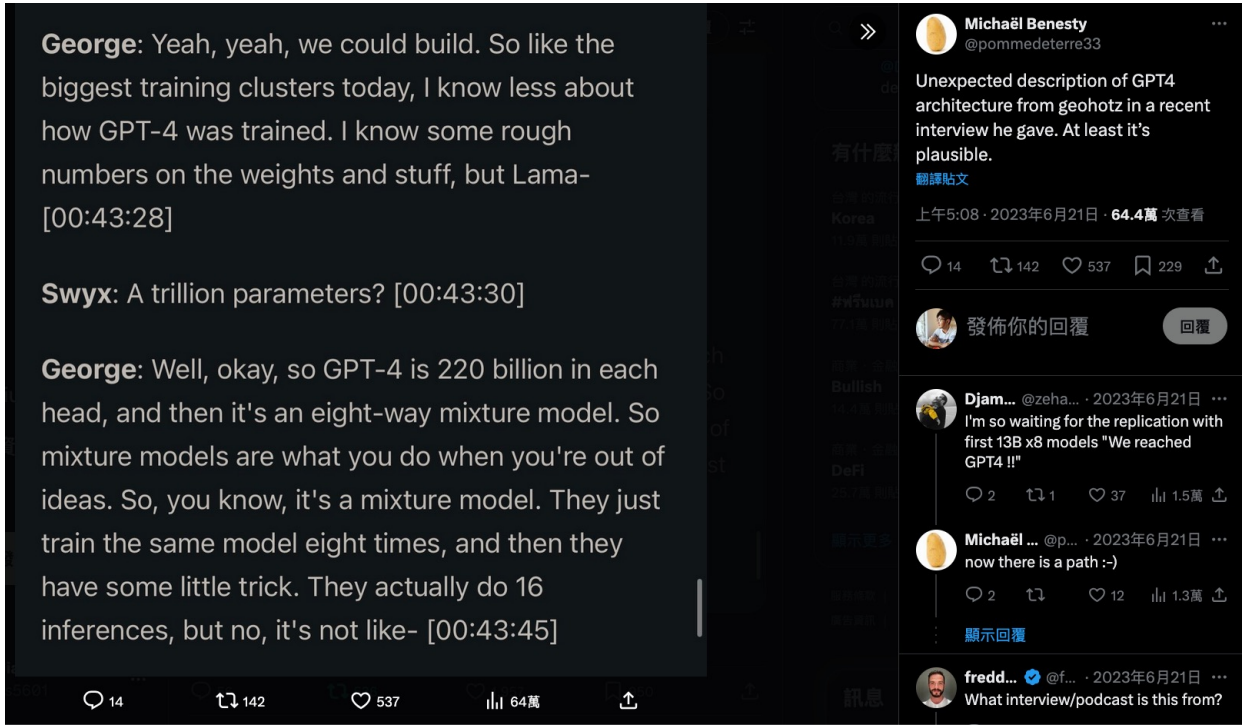


Why do we need to learn MoE? (GPT-4)



PyTorch 創始者推測 GPT-4 採用 MoE

https://x.com/soumithchintala/status/1671267150101721090



Why do we need to learn MoE? (Popular LLMs)

Model	Release Date	Active Parameters (Total Parameters)	Company
DeepSeek-V3	2024/12/27	37B (671B), 256 experts	DeepSeek
DeepSeek-R1	2025/1/22	37B (671B), 256 experts	DeepSeek
Llama 4 Maverick	2025/4/5	17B (400B), 128 experts	Meta
Mixtral 8x7B	2024/1/8	13B (47B), 8 experts	Mistral AI

Total #models: **243**. Total #votes: **2,945,410**. Last updated: 2025-05-22.

Code to recreate leaderboard tables and plots in this notebook. You can contribute your vote at lmarena.ai!

Category: Overall

Apply filter: Style Control | Show Deprecated

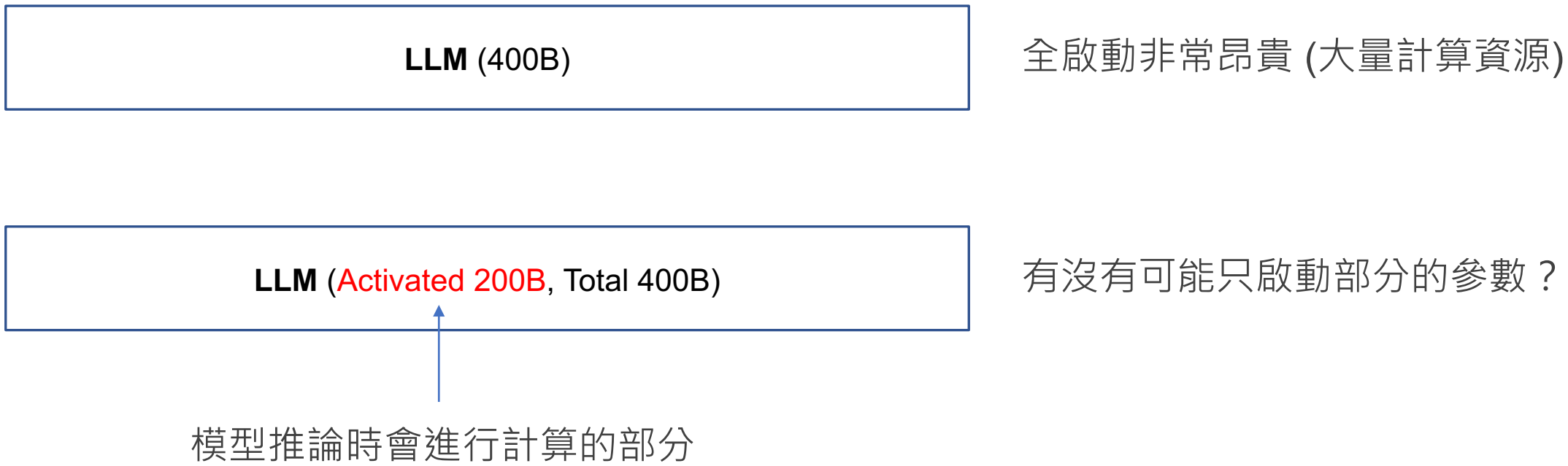
Overall Questions

#models: **243 (100%)** #votes: **2,945,410 (100%)**

Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score	95% CI	Votes	Organization	License	Knowledge Cutoff
1	1	Gemini-2.5-Pro-Preview-05-06	1446	+6/-7	6115	Google	Proprietary	Unknown
2	3	Gemini-2.5-Flash-Preview-05-20	1418	+10/-10	3892	Google	Proprietary	Unknown
2	1	o3-2025-04-16	1409	+7/-6	7921	OpenAI	Proprietary	Unknown
2	2	ChatGPT-4o-latest (2025-03-26)	1405	+6/-5	10280	OpenAI	Proprietary	Unknown
3	6	Grok-3-Preview-02-24	1399	+5/-3	14840	xAI	Proprietary	Unknown
4	3	GPT-4.5-Preview	1394	+5/-4	15276	OpenAI	Proprietary	Unknown
6	6	Gemini-2.5-Flash-Preview-04-17	1387	+7/-8	6938	Google	Proprietary	Unknown
8	6	DeepSeek-V3-0324	1368	+5/-5	9741	DeepSeek	MIT	Unknown
8	6	GPT-4.1-2025-04-14	1365	+8/-8	6094	OpenAI	Proprietary	Unknown
8	13	Hunyuan-Turbos-20250416	1356	+9/-7	5111	Tencent	Proprietary	Unknown
9	9	DeepSeek-R1	1354	+4/-4	19339	DeepSeek	MIT	Unknown
10	18	Gemini-2.0-Flash-001	1351	+4/-3	24928	Google	Proprietary	Unknown
10	13	Mistral Medium 3	1343	+11/-10	3327	Mistral	Proprietary	Unknown

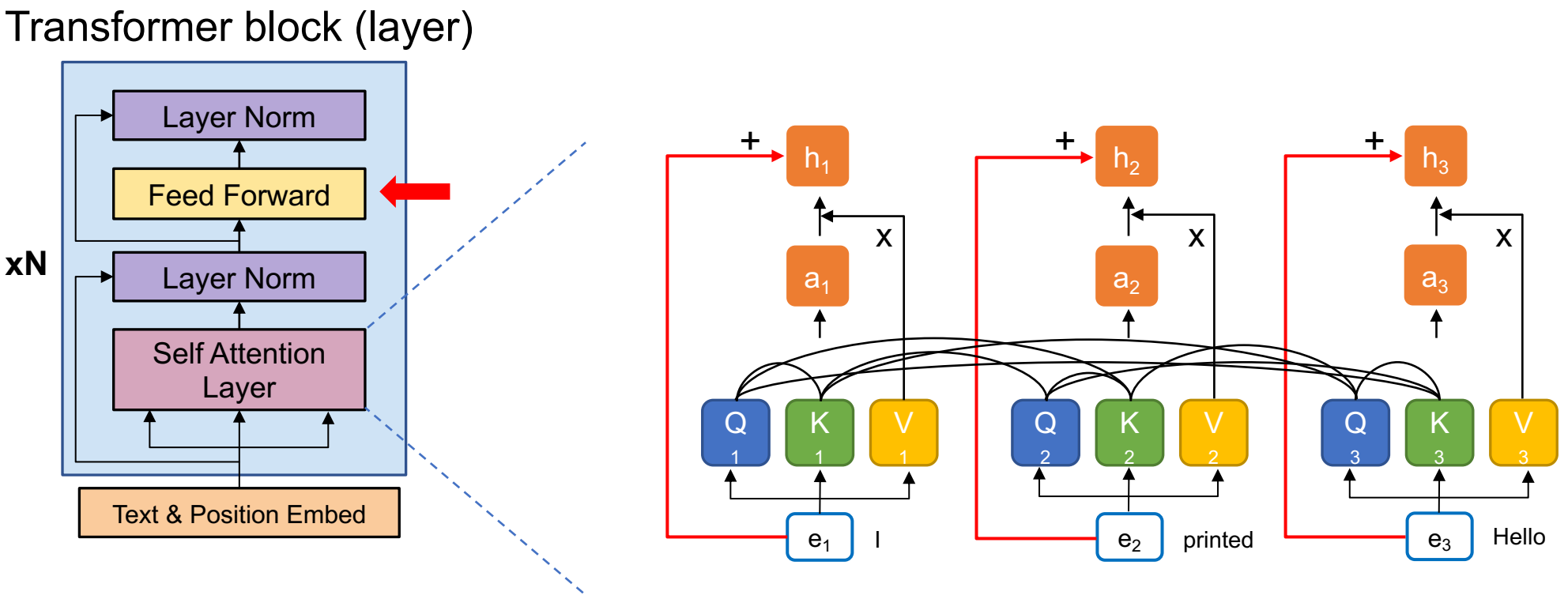
https://huggingface.co/spaces/lmarena-ai/chatbot-arena-leaderboard

Why do we need ~~to learn~~ MoE?

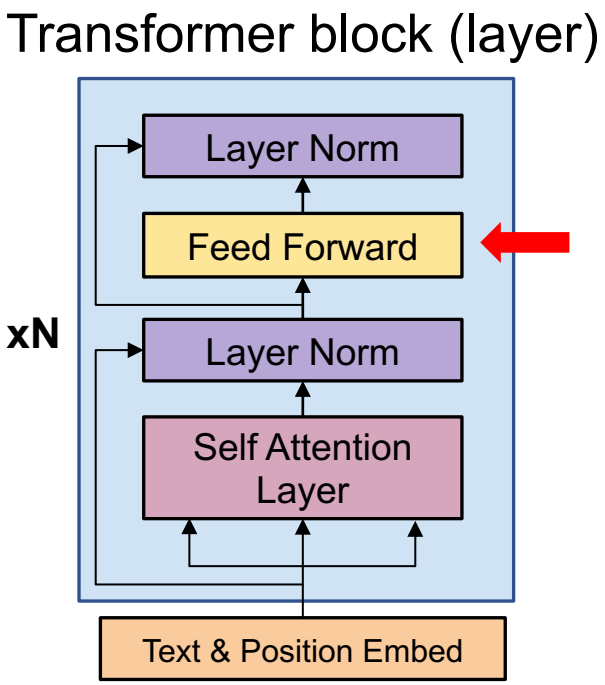


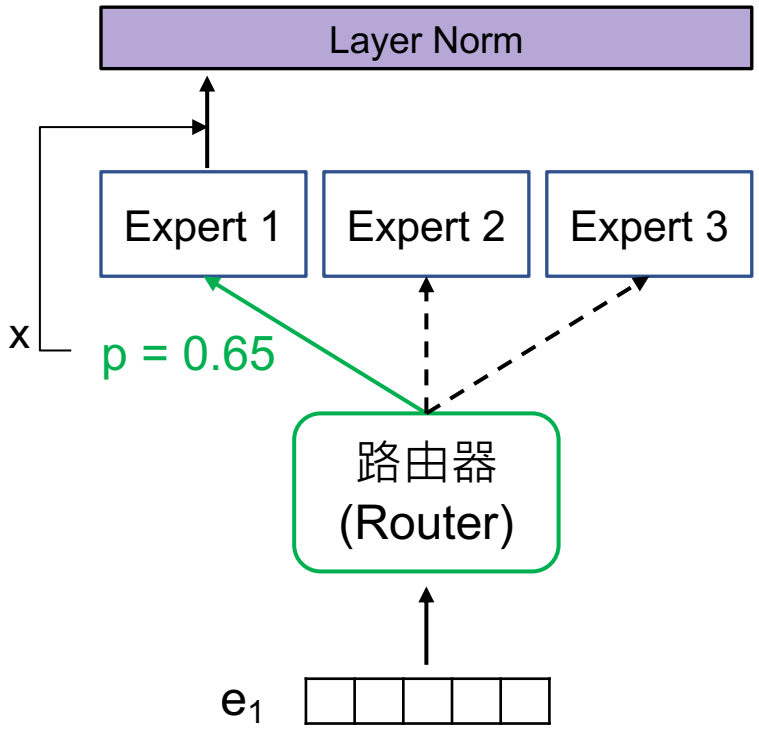
MoE

[Recap] Transformer block

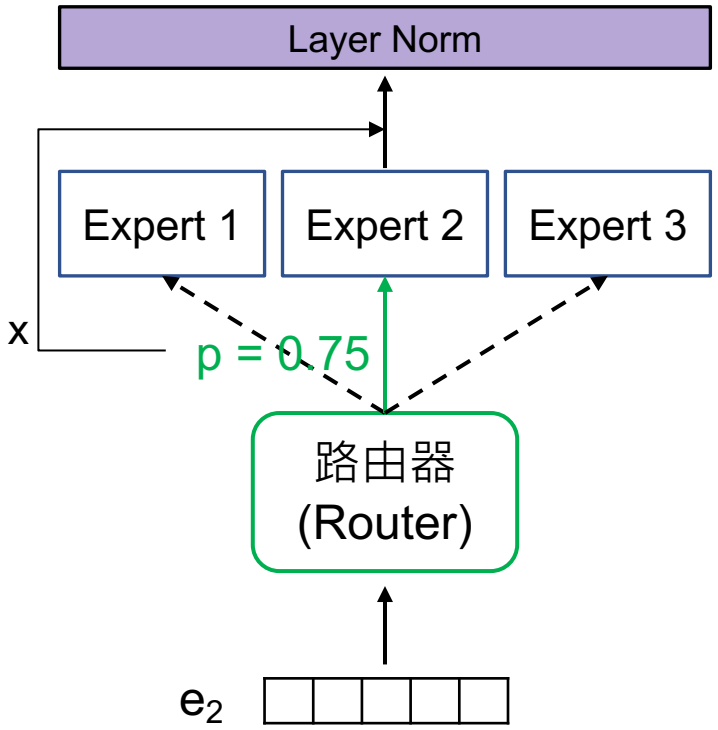


Mixture of Experts (MoE)





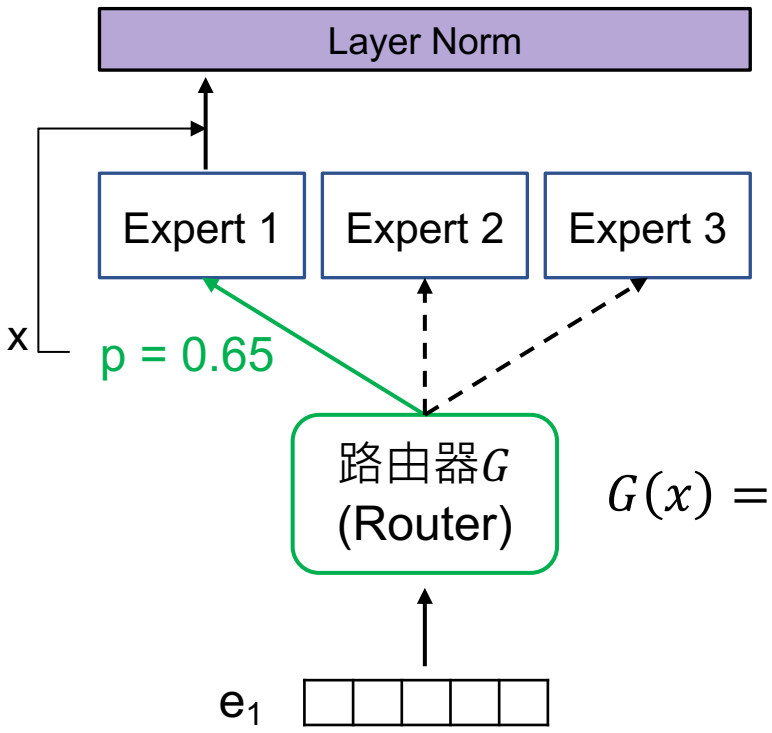
(來自第一次 Add & Norm 的輸出)



(來自第一次 Add & Norm 的輸出)

Router

- Router 是一個小型神經網路，用來決定輸入 token 該被送到哪一個或哪幾個 experts，又稱作 Gating Network
- 每個 token 都會依照機率來選擇 Top-k 個 experts



$$G(x) = \text{Softmax}(x \cdot W_g)$$

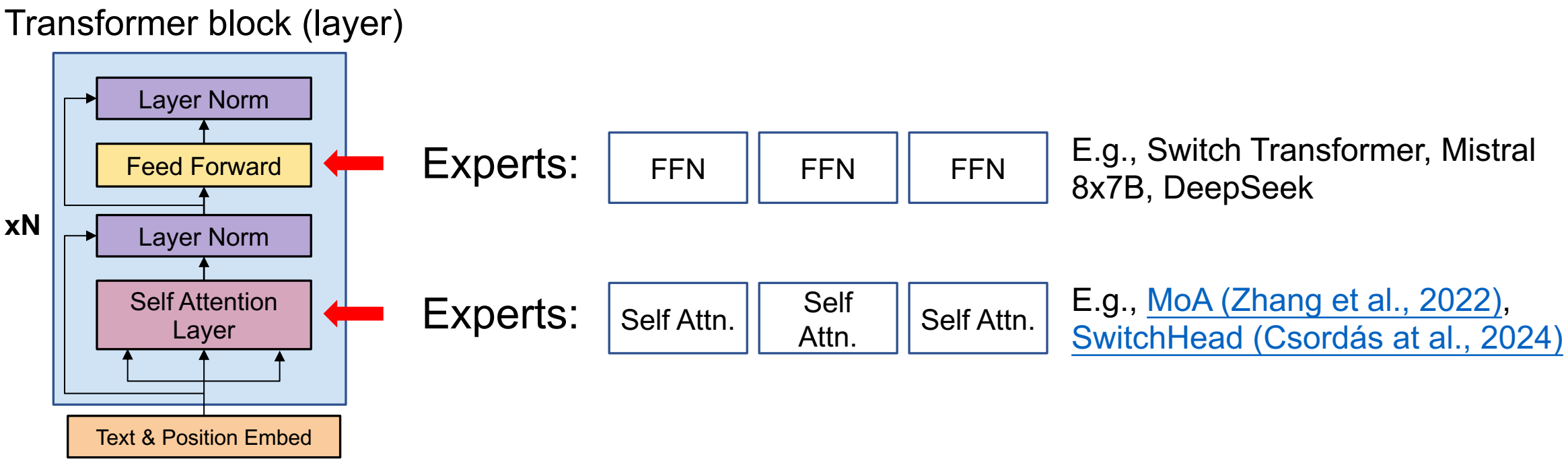
$W_g \in \mathbb{R}^{d \times E}$

d: x 的向量維度

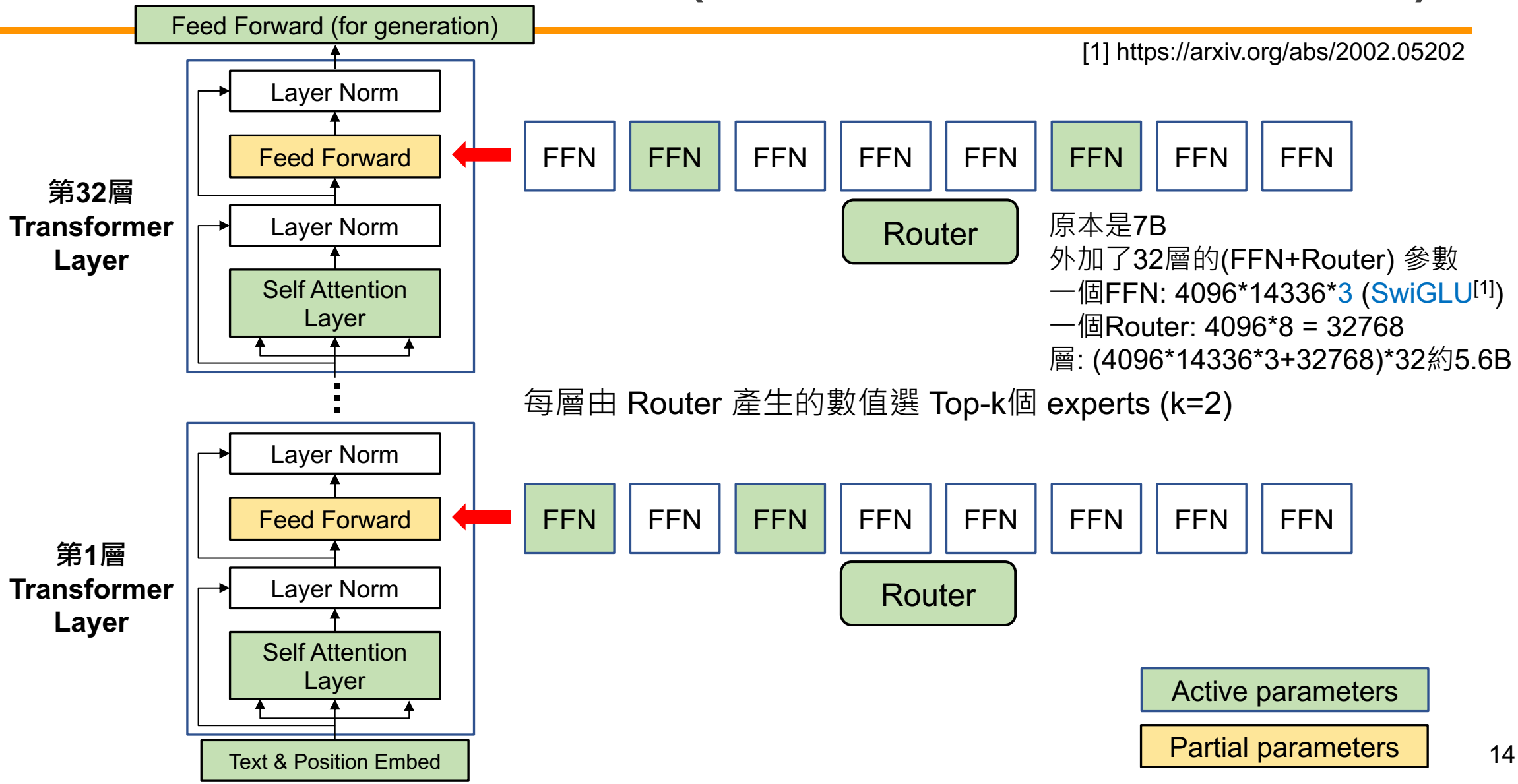
E: experts 的數量

Expert 是什麼？

- 主要是看你放在哪裡，Expert 就會是一樣的架構



Active Parameters (以Mistral 8x7B為例)



[1] https://arxiv.org/abs/2002.05202

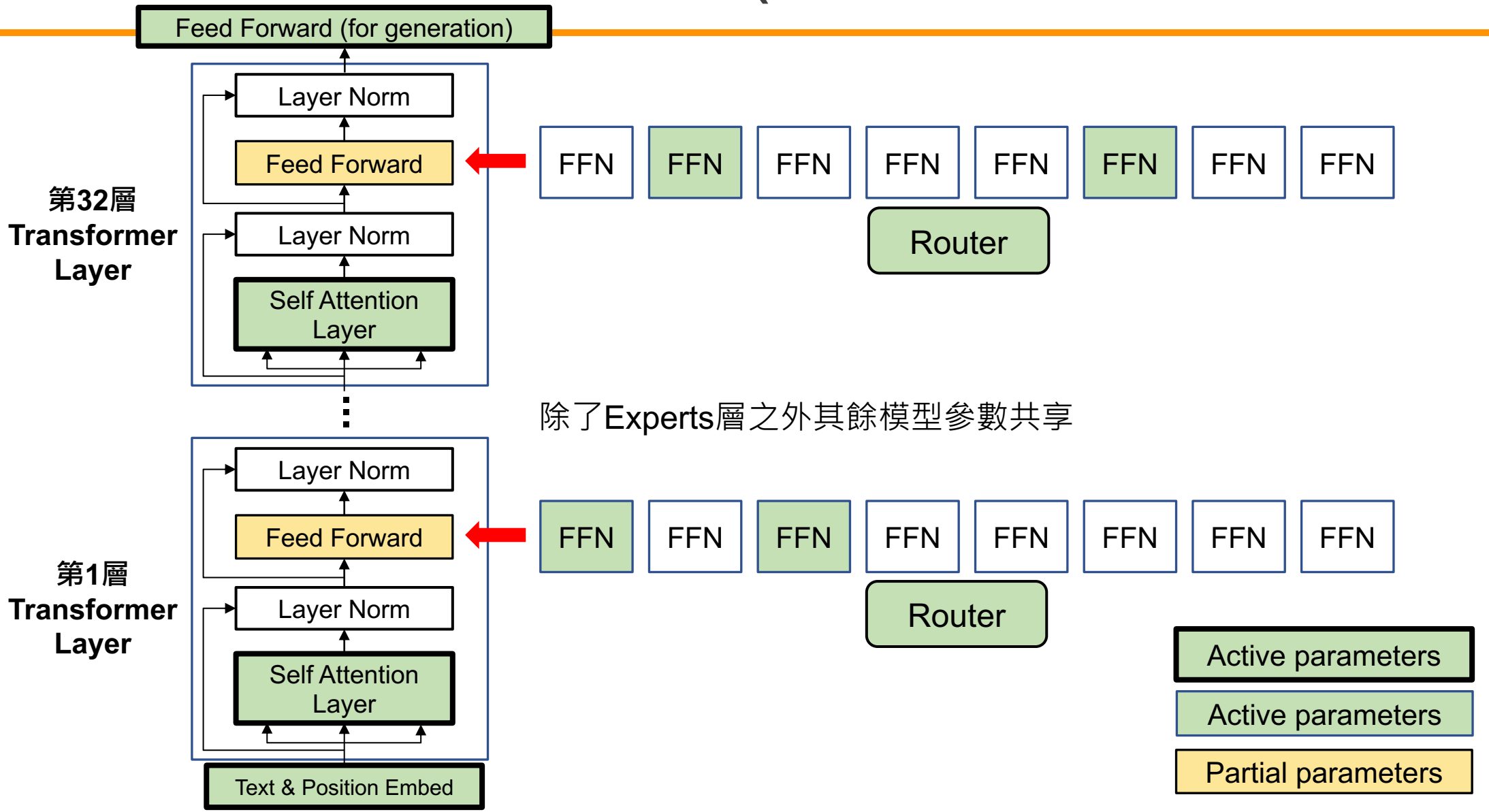
原本是7B
外加了32層的(FFN+Router) 參數
一個FFN: 4096*14336*3 (SwiGLU[1])
一個Router: 4096*8 = 32768
層: (4096*14336*3+32768)*32約5.6B

每層由 Router 產生的數值選 Top-k個 experts (k=2)

Active parameters

Partial parameters

Shared Parameters (以Mistral 8x7B為例)



Take Home Message

- 現代 LLM 越來越大，全參數啟動非常昂貴
- 一般模型 vs. MoE
  - 一般模型在推論階段會同時啟用所有參數，即使部分輸入根本不需要那麼多處理能力
  - MoE 採用 conditional computation：根據輸入內容，只啟動部分 experts，因此可以減少記憶體使用量，加快模型推論速度
- MoE 額外好處：Experts 數量可以增加，但 activated parameters 數量可以維持

MoE vs. Ensemble Learning

- 推論時使用參數量比較：
  - MoE: 少數專家 (sparse)；Ensemble: 各模型全部
- 參數共享：
  - MoE: 專家層外參數共享；Ensemble: 無參數共享
- 訓練方式：
  - MoE: 所有專家一起訓練；Ensemble: 不同模型可以各自訓練

Online resources

- YouTube videos
  - Stanford CS25: V4 I Demystifying Mixtral of Experts
  - [IBM Technology] What is Mixture of Experts?
  - A Review of 10 Most Popular Activation Functions in Neural Networks
- Important papers
  - 近代 MoE 開山之作
    - Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer (Shazeer et al., 2017)
  - MoE on Transformers (T5)
    - Switch Transformers (Fedus et al., 2021)

Thank you!

Instructor: 林英嘉

✉ yjlin@cgu.edu.tw