政大金融AI雲-
Training Computation-Optimal Finance Domain LLMs

Rua-Huan Tsaih and **Fang Yu**

Dept. Management Information Systems
National Chengchi University

AMD AI Solutions Day 11/5/2024

NCCU + AMD

政大與AMD攜手合作　建立臺灣北部AI生態圈

日期：2024-10-01　單位：秘書處

【校訊記者許巧昕報導】

國立政治大學與美商超微半導體公司（以下稱AMD）強強聯手，日前共同簽訂產學合作備忘錄實力，積極推動臺灣北部AI生態圈之建立，著重於培養具備AI應用能力的人文社會科學領域專

痛點和動機

- 大型語言模型（LLMs）是人工智慧（AI）領域中的關鍵技術，能夠處理多種自然語言任務，如文本生成、問答系統及摘要等。
- 在金融領域，金融LLMs具有巨大的潛力，尤其是在風險管理、財務分析及智能投資等應用場景。
- 然而，通用LLMs由於缺乏特定金融領域的深入知識，難以準確理解專業術語及複雜的財務概念，導致其在面對專業性強、精度要求高的金融任務時，存在重大挑戰。例如，通用LLMs在處理財務報告解讀、合規分析及市場預測時表現不佳。
- 建議解決方案：金融領域腦 (Finance Domain LLMs)
- That is: 利用合適之持續預訓練（Continual Pretraining）技術以及持續收集之專業金融領域語料，使金融領域腦在金融應用中，可以克服金融知識淺層的問題，更具專業性及準確性，同時還能定期更新模型知識，確保其應用不落後於快速變化的金融市場。

文獻回顧：台灣繁中腦

113/4/19 Meta 公布 Llama3，TAIDE團隊僅花四天時間完成模型之訓練及經過基本驗測，並循程序獲得國科會同意後於公開釋出以Llama3為基底的Llama 3-TAIDE-LX-8B-Chat-Alpha1模型搶先版。

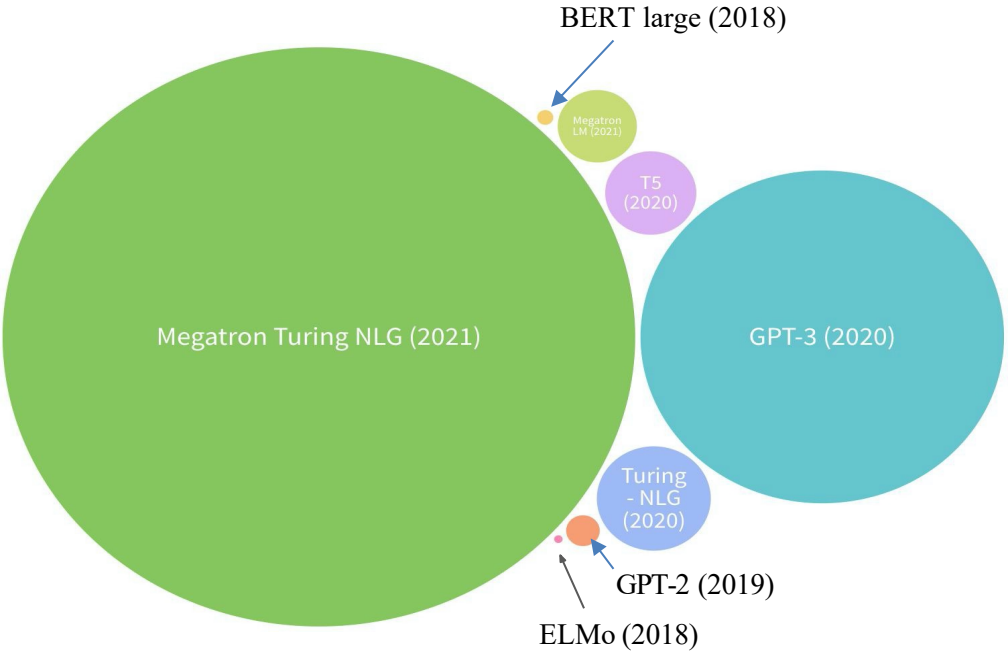
113/5/3 TAIDE計畫歷經一年，開發出基於Llama2的可商用TAIDE LX-7B模型、可學研用的TAIDE LX-13B模型，以及可商用的Llama 3-TAIDE-LX-8B-Chat-Alpha1等三種版本，特舉辦成果發表會，並邀請合作夥伴展示各種TAIDE應用成果。

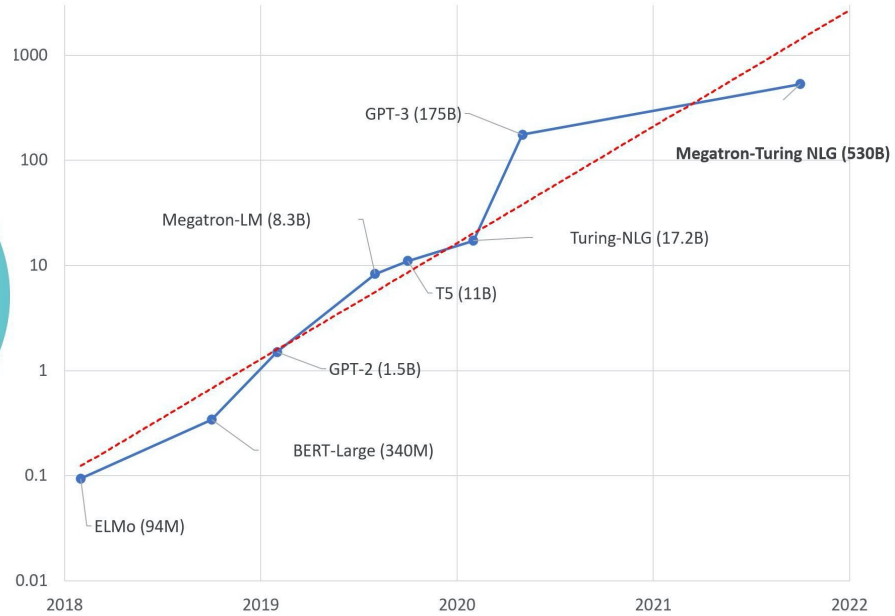
產生金融領域腦之做法

比照台灣繁中腦(TAIDE or Breeze)之作法：用足夠量(x_1 B tokens)的優質繁體中文語料集持續預訓練(continual pre-training) Llama 3 成 台灣繁中腦

1. 用足夠量(x_2 B tokens)的優質繁體中文金融語料集持續預訓練 台灣繁中腦 成 台灣金融腦 – 繁中金融專業資料強度為$\frac{x_2}{x_1+x_2}$
2. 用足夠量(x_3 B tokens)的A銀行優質繁體中文金融語料集持續預訓練 台灣金融腦 成 台灣A銀行腦 – 繁中A銀行金融專業資料強度為$\frac{x_2+x_3}{x_1+x_2+x_3}$
3. 用足夠量(x_4 B tokens)的A銀行優質繁體中文信評業務語料集優調(fine-tuning) 台灣金融腦 成 台灣A銀行信評腦 – 繁中A銀行信評金融專業資料強度為 $\frac{x_2+x_3+x_4}{x_1+x_2+x_3+x_4}$

Language Models are Getting Bigger…





[Image Source] [Image Source]

Understanding FLOPs

$$\mathbf{C \sim 6ND}$$

If we had a **computational budget** on C,
Increasing model size **N** = **Decreasing** dataset size **D**

But we also expect **more data → better performance**

C = number of FLOPs (computations)
N = number of model parameters
D = amount of training data

Key Question

*To **maximize** model performance,*

*how should we **allocate C** to **N** and **D**?*

$$N_{opt}(C), D_{opt}(C) = \underset{N,D \text{ s.t. } \text{FLOPs}(N,D)=C}{\text{argmin}} L(N, D)$$

[Equation Source: (Hoffman et al., 2022)]

Key Question (rephrased)

*What is the **relationship** between **loss** and **N, D**?*

$$\hat{L}(N, D) \triangleq E + \frac{A}{N^{\alpha}} + \frac{B}{D^{\beta}}$$

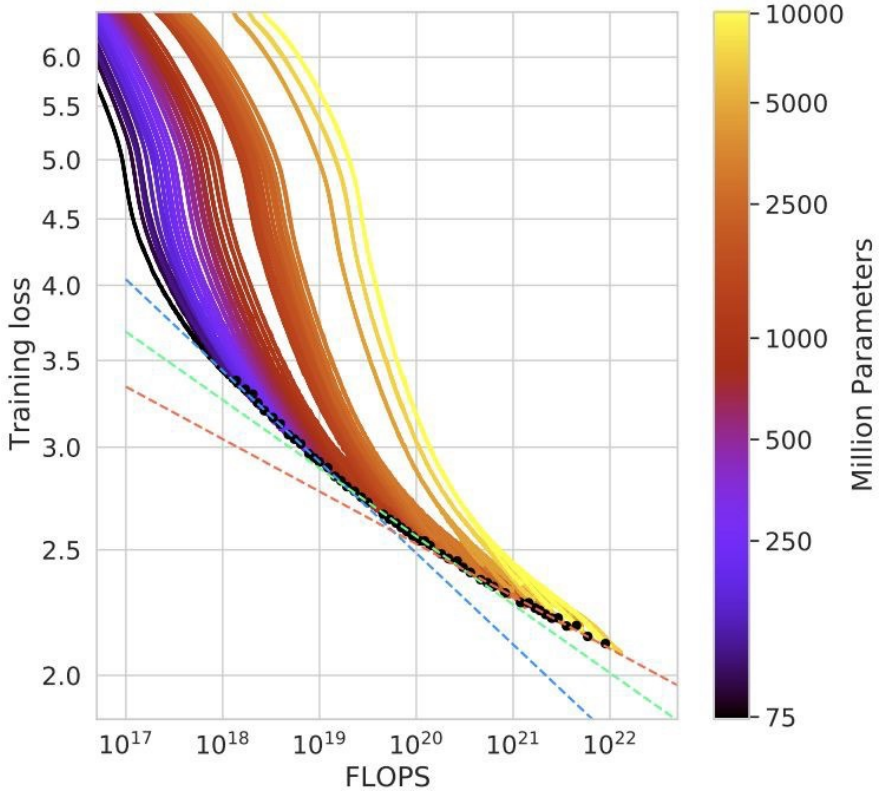
[Equation Source: (Hoffman et al., 2022)]

Is Power-Law the best fit?

Based on **empirical observation**

No theoretical background

(Hoffman et al.) also observe concavity in their model at high compute budgets, suggesting the **need for a more detailed model**



[Figure Source: (Hoffman et al., 2022)]

Data Pruning (Sorcher et al., 2022)

Develop a metric to measure the **quality of data**

Prune the data to include only high quality data

Importance of dataset size **decreases** significantly

Web/Open Corpus

數據類型	數據格式	總量	描述	總 Tokens
金融、證券、保險博碩論文	PDF 轉 JSON	70頁 × 210篇	涵蓋多個金融相關主題的博碩論文	3,087,000
金融、證券、保險教科書	PDF 轉 JSON	18,662頁	大型金融、證券、保險教科書，結構化章節內容	9,331,000
金融、證券、保險法規及開放數據	HTML 轉 JSON	15,898,700字（繁體中文）	涵蓋金融法規、開放數據等結構化條款	31,797,400
金融保險試題	PDF 轉 JSON	7.18M tokens	涵蓋多種考點的金融保險相關試題	7,180,000
金融保險問答	JSON	0.89M tokens	涵蓋常見問題的金融保險相關問答資料	890,000
FinGPT 數據集	JSON	16.7M tokens	為 FinGPT 設計的金融語料資料集	16,700,000
UltraChat（經過關鍵詞過濾）	JSON	970M tokens	經關鍵詞過濾的聊天記錄，涵蓋金融相關對話	970,000,000
Common-crawl-zhtw	JSON	1.55B tokens	繁體中文網頁資料，涵蓋多領域內容	1,550,000,000
CC-100-zh-Hant-mertges	JSON	5.11B tokens	繁體中文網頁資料，包含多語種與多領域內容	5,110,000,000
c4-zhtw	JSON	1.1B tokens	繁體中文網頁資料，來自 C4 語料庫	1,100,000,000
zhtw-news and article-2B	JSON	1.47B tokens	繁體中文新聞及文章資料	1,470,000,000
wikipedia-zhtw-dedup	JSON	0.75B tokens	繁體中文維基百科去重資料	750,000,000

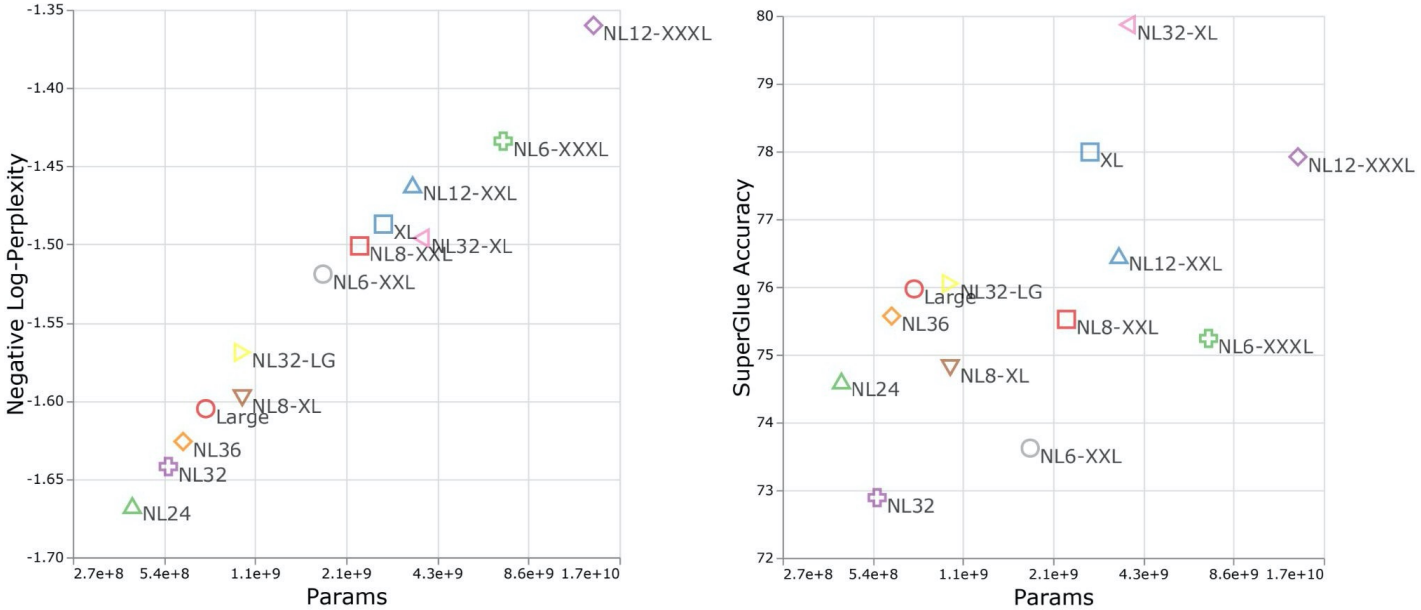
共 11,018,985,400 tokens

Scaling Law For Fine-Tuning (Tay et al., 2021)

Downstream performance **after fine-tuning** does not scale with model size

Downstream performance does **scale with depth**, but not necessarily with dimension

Downstream Performance Does Not Depend on N



(a) Pre-training scaling

(b) Fine-tuning scaling

[Figure Source: (Tay et al., 2021)]

Training all v.s. specific layers

With limited dataset, training a model with a large number of parameters may lead to overfitting.

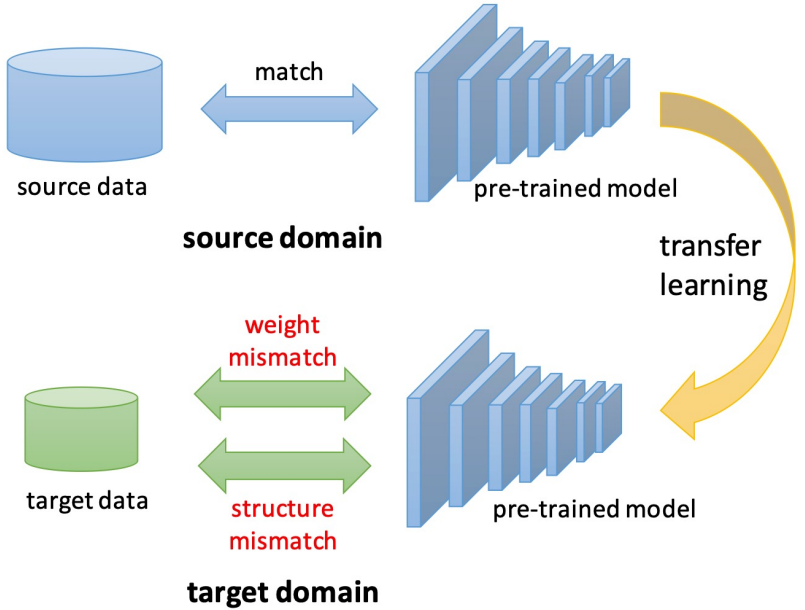


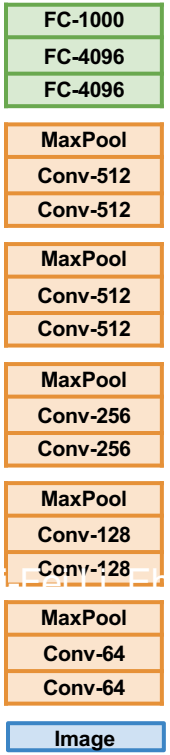
Figure 1: Illustration of the two mismatches during transfer learning.

Reference: TransTailor: Pruning the Pre-trained Model for Improved Transfer Learning

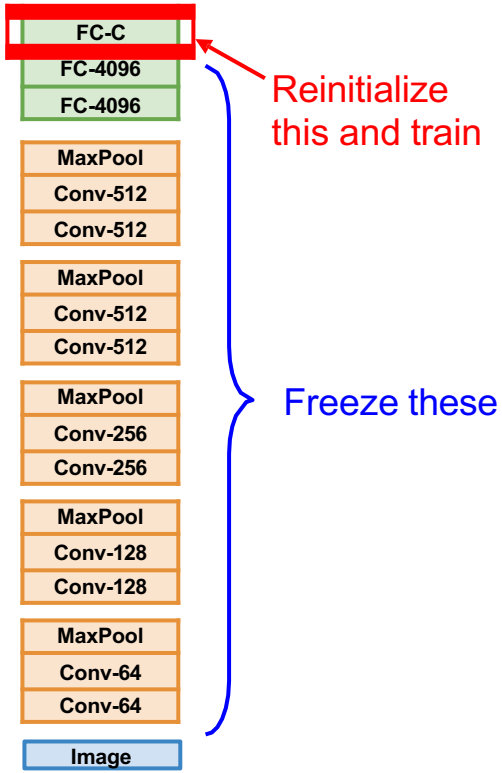
Transfer Learning with CNNs

Donahue et al, "DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition", ICML 2014
Razavian et al, "CNN Features Off-the-Shelf: An Astounding Baseline for Recognition", CVPR Workshops 2014

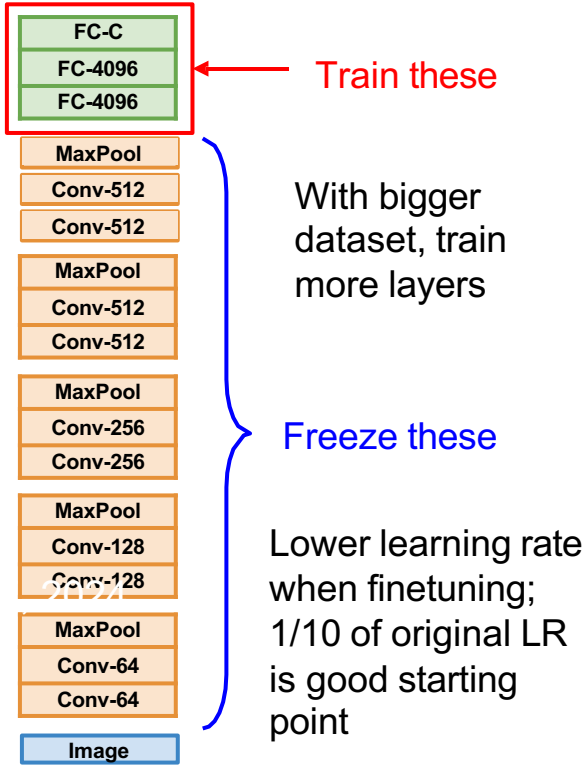
1. Train on Imagenet



2. Small Dataset (C classes)

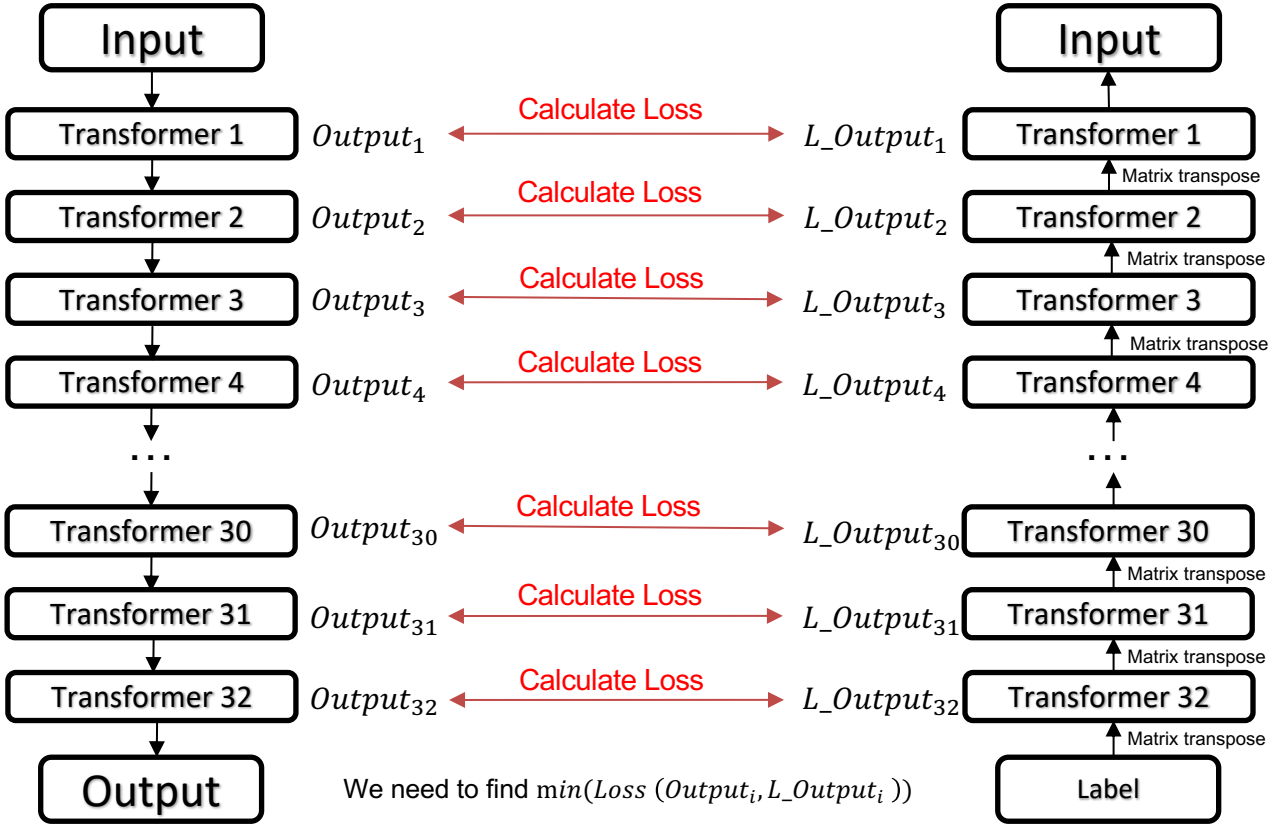


3. Bigger dataset



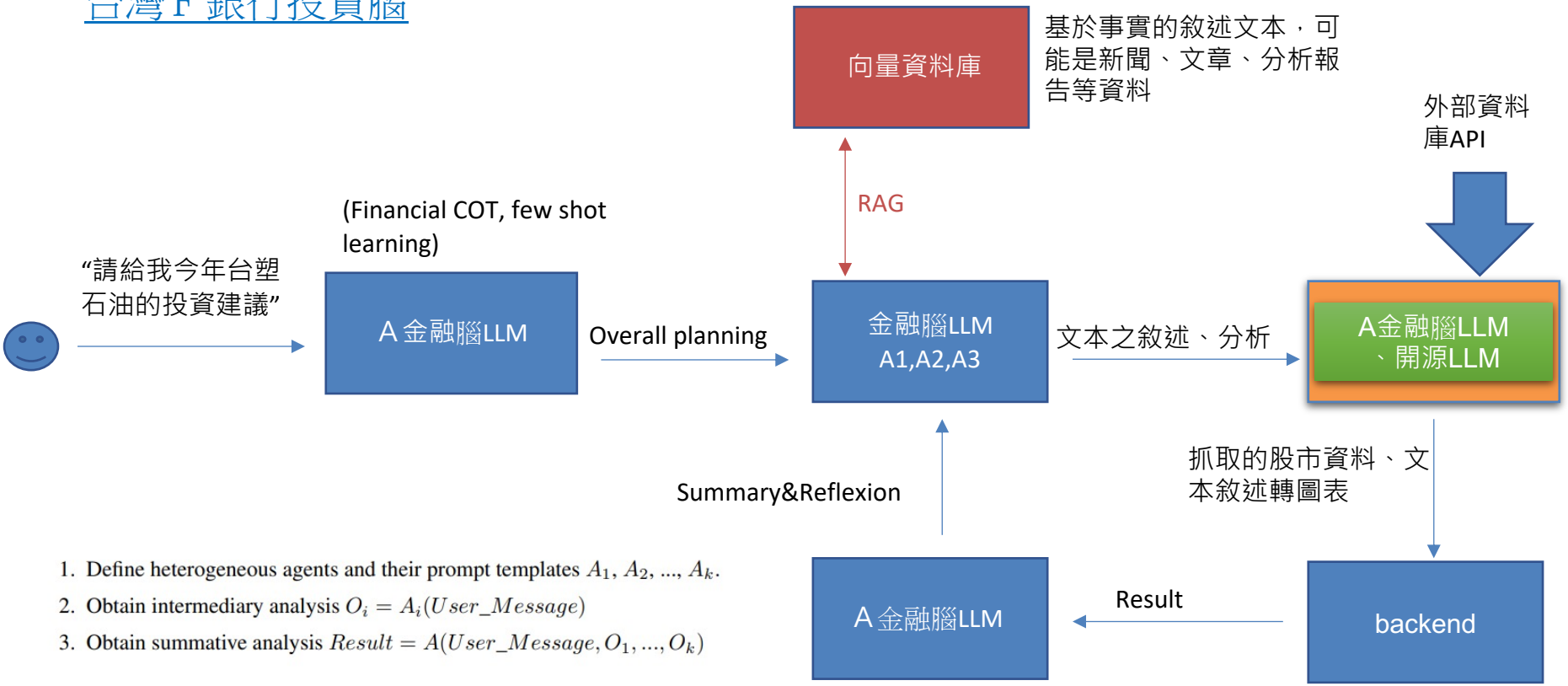
How to select layers to be trained?

Reference:LogME: Practical Assessment of Pre-trained Models for Transfer Learning (2024)



We need to find $\min(Loss\ (Output_i, L\_Output_i))$

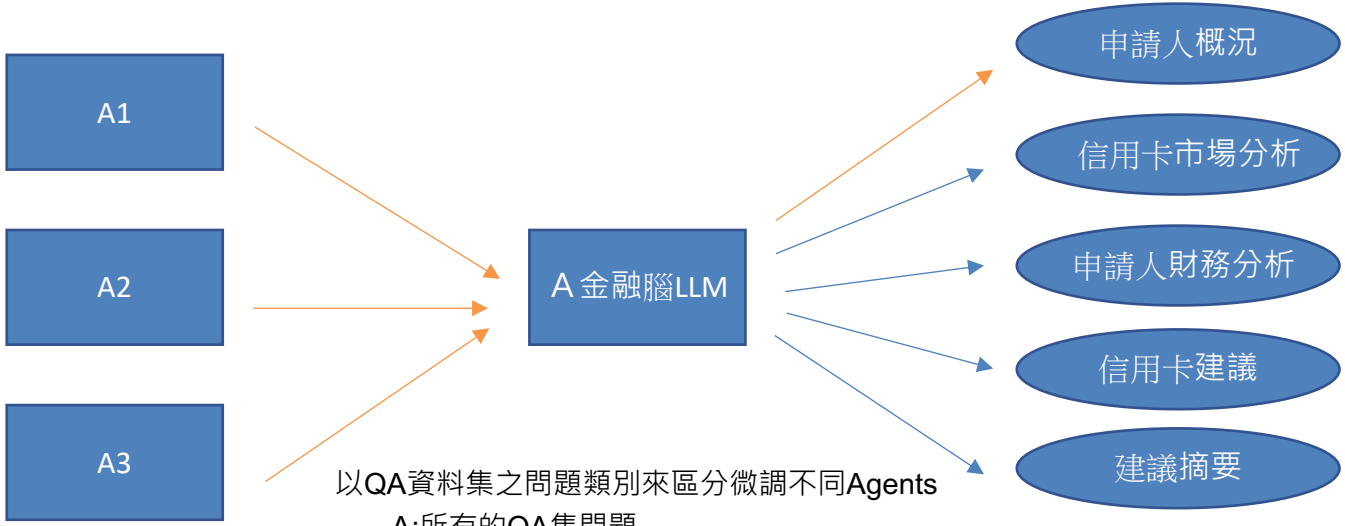
台灣F銀行投資腦



1. Define heterogeneous agents and their prompt templates $A_1, A_2, ..., A_k$.
2. Obtain intermediary analysis $O_i = A_i(User\_Message)$
3. Obtain summative analysis $Result = A(User\_Message, O_1, ..., O_k)$

台灣C銀行信評腦

期望透過HAD框架來去提高效能，並加入RAG避免幻覺，提高問答準確率



以QA資料集之問題類別來區分微調不同Agents

A:所有的QA集問題
A1:信用卡和申請人相關資料
A2:信評和風評
A3:法規

最後將A1,A2,A3之答案輸入A，共同輸出彙總

Recap: 政大金融腦

- Collect high-quality finance corpus
- Derive proper continual pretraining algorithm and fine-tuning algorithm
- Explore the applications
- Thanks again for the support of AMD.

政大攜手AMD將開AI課 打造人文導向研究中心

2024-10-01 16:18 中央社／台北1日電

+ 公共政策

分享 0 分享

政治大學與美商超微半導體公司（AMD）簽訂產學合作備忘錄，雙方將共同開設人文社會科學導向的AI賦能課程，並打造人文社科為核心的AI研究中心，推動台灣北部AI生態圈的建立。

政治大學今天發布新聞稿指出，政大以人文社會科學為核心優勢，過去不斷探索如何將人文社科領域與現代科技相結合；此次與AMD合作，象徵學術機構與科技產業的深度融合，雙方共同期望通過此次合作，提升政大在AI研究的實力，也讓AI在人文社科領域有更廣泛的應用。

政治大學 NATIONAL CHENGCHI UNIVERSITY