



台灣大哥大

AI 2.0
TECHForum

線上技術論壇

BYOM-打造企業專屬的大語言模型及應用

Andy Huang 黃世民
台智雲 AI算力事業發展處

2024/03/14

生成式 AI 正在改變我們的世界





不只個人應用 生成式 AI 更為組織帶來加乘效果

80% 的受訪者表示
透過生成式 AI 可加速工作效率

40% 以上的受訪者認為
生成式 AI 可增加跨界或斜槓的競爭力

90% 以上的管理職考慮或已經
導入生成式 AI 工具於公司營運中

AI 2.0 – 各行各業的應用爆發

媒體雜誌 (文本生成、知識庫搜尋)

- 過去累積大量文本資料(專刊、報導、短訊..)，經 Finw Tune 創生集團企業腦，加速文本生成速度及數量 (Client Capability)
- 透過 Embedding 架構，建立專業知識庫應用 (Chat KM)
- 財經、健康、生技.....等應用

醫療健康機構 (文本生成、線上健康諮詢)

- 快速盤點各項健檢資訊重點，生成客戶端摘要報告說明
- 透過 Line Chat 客服累積專業健康知識對話資料，建立自動化 Chat 健康諮詢智能對話機制

服務 / 零售 (精準行銷推薦、行銷文案)

- 透過過去大量累積銷售結構性資料，進行客戶精準推薦機制
- 透過銷售KM資料，快速生成客戶行銷商品推薦建議

電玩遊戲 (虛擬人物應用)

- 透過故事情境文本，創建 NPC 虛擬智慧人物場景

傳統製造 (文件規格辨識、重點摘要產出)

- 制式化文件重點關鍵字義 辨識及相關重點摘要生成
- 應用在 規格書、工程技術說明、標單文件.....

科技製造 (產線資料分析應用、企業內部智能輔助)

- 整合 AI 1.0，建立智慧工廠產線資料分析及對應作業摘要生成
- 大量會議紀錄重點摘要生成、翻譯工作、KM 知識庫搜尋

電信商 (線上客服)

- 透過客服 QA / QD 資料庫，建立自動化擬人客服智能對話機制
- 快速累積對談資料集，持續進行 優化企業腦

金融金控 (生成式AI數位架構作業)

- 建立企業專屬生成式AI數位架構作業機制(環境)，以因應企業集團多樣化終端應用服務
- 法尊稽核程序合規：企業腦雲端微調作業、模型地端派送部署、營運作業稽核、營業秘密保障、資訊安全作業
- 持續性營運預算合理性支持 (訓練創建 + 營運應用)

市場趨勢: 生成式AI 帶來巨大的生產力革命

【2023.06.14 諮詢機構麥肯錫】

《生成式人工智慧的經濟潛力》研究報告，分析師們通過對全球47 個國家（全球80%以上勞動人口）、60多個組織、約850個職業、2100項工作活動的勞動生產力潛力的研究，探討在 AI 成指數級發展背後，對全球經濟將帶來的影響。

<https://www.blocktempo.com/mckinsey-research-report-on-generative-ai/>

AI 取代人類工作的時間被大幅提前了 10 年，在 2030 年至 2060 年間（中點為 2045 年）50% 的職業逐步被 AI 取代

AI 每年可為全球經濟帶來 2.6 兆至 4.4 兆美元的增長，生產力提高 0.1%—0.6%，相當於每年貢獻一個英國的 GDP

全域性上看 AI 對各行各業的發展有利，但是對個人不利，而高薪、高學歷的腦力勞動者受到的衝擊最大

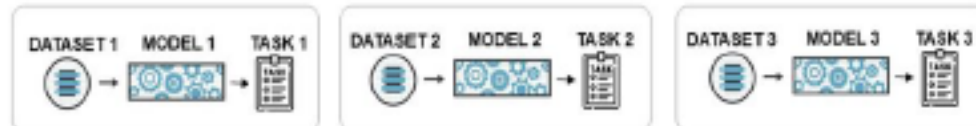
生成式 AI 帶來的價值增長，主要（約 75%）集中在四個領域：客戶運營、行銷和銷售、軟體工程和研發

生成式人 AI 及其他科技的發展或將使當前工作的 60% 到 70% 實現自動化。銀行業、高科技和生命科學等影響最大

AI 2.0 克服了 AI 1.0 單領域、多模型得限制（引用 創新工廠）

單一領域数据集：诸多数据集和诸多模型各成孤岛缺乏纵效；劳动密集型的数据标注

AI 1.0

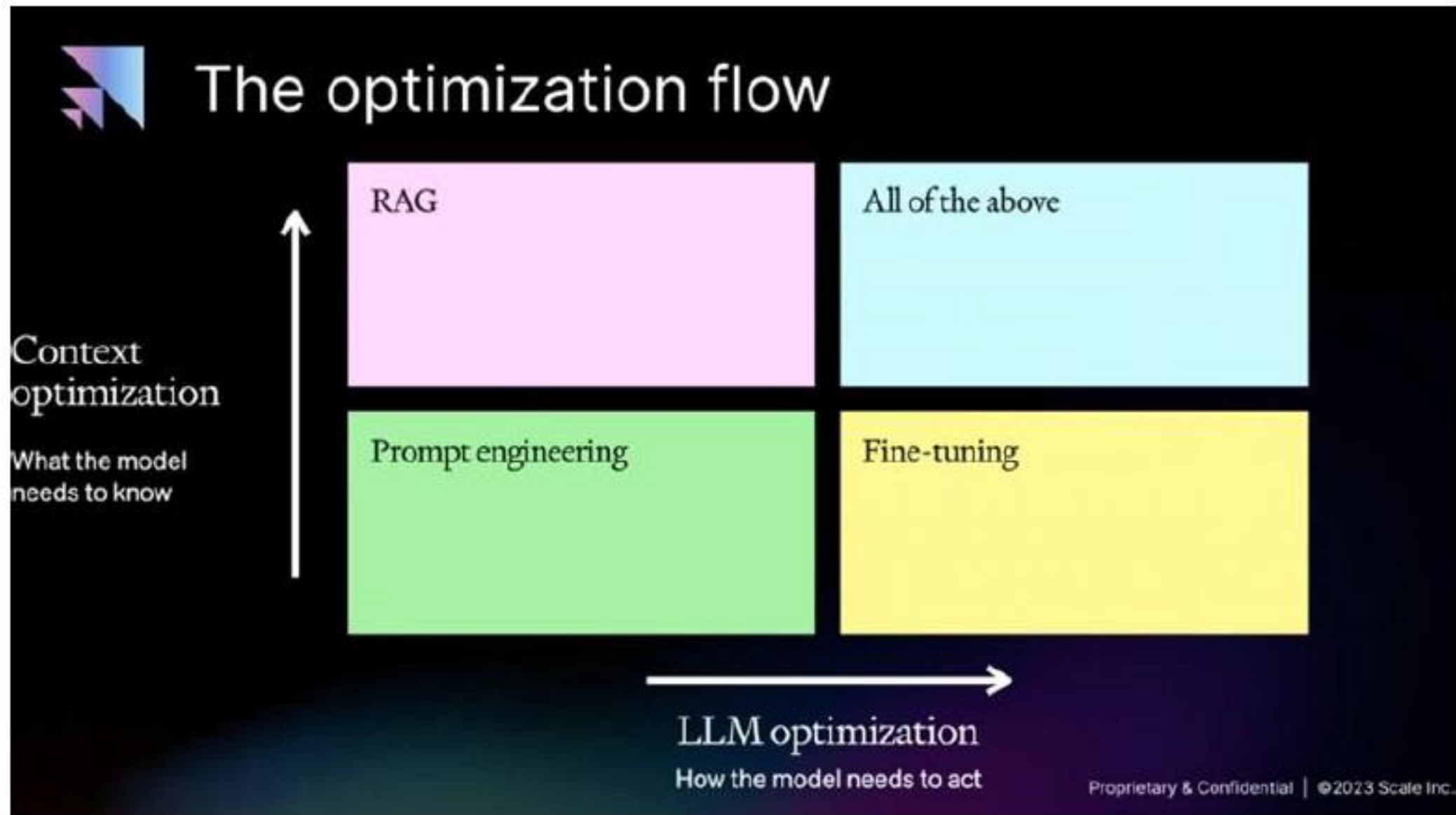


超级海量数据，无需人工标注；一个具有跨领域知识的“基础模型”，能执行五花八门的任务

AI 2.0



訓練企業大腦的四種方法



第二大腦：以 RAG 技術打造企業專屬 AI 雲端知識庫



消弭 AI 知識幻覺

- 透過LLM模型提升語意理解
- 語意搜索精確擷取需求資訊
- 知識庫隨時更新保障來源



降低導入及應用成本

- 降低生成技術的佈署成本
- 採預訓練模型降低導入門檻
- 加速應用，實現高效流程



自主選擇資料環境

- 符合組織內特定權限及政策
- 滿足隱私及合規性要求
- 實現更大的靈活性及安全性

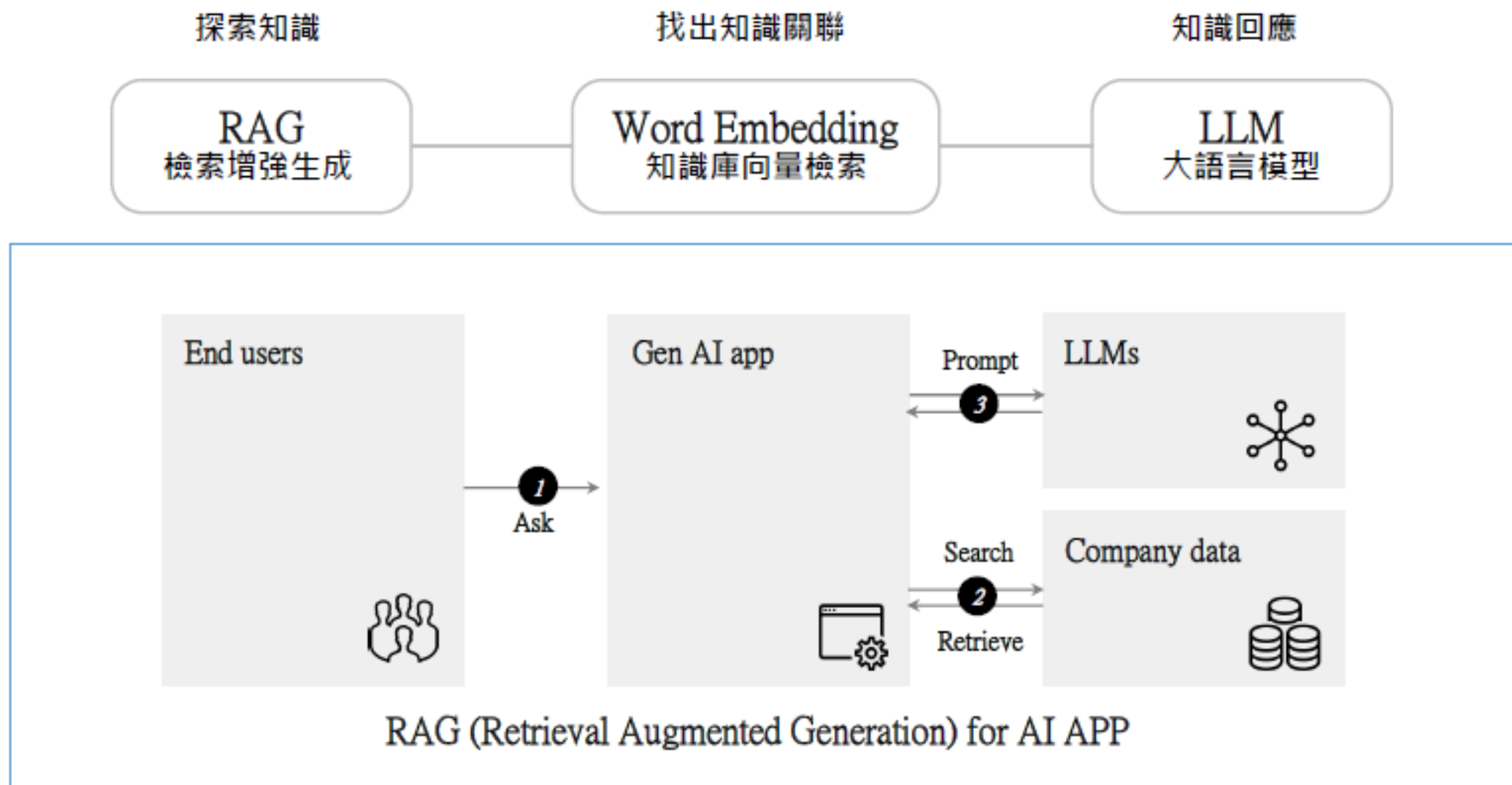
知識管理
Knowledge Management

- 自動摘要生成，輔助節省工作成本
- 知識庫問答系統，快速回應提問
- 知識分享平台，員工輕鬆找尋資訊
- 多語言翻譯和生成，跨語言溝通支援
- 快速打造專業領域知識庫，如法規政策

知識服務
Knowledge Service

- 智能客服，解答民眾常見提問
- 虛擬助手，協助執行特定內容任務
- 教育培訓，輔助教育或專業人士累積知識
- 民意蒐集，提供長官調整施政相關策略

RAG 檢索增強生成技術的原理



Coding Assistant

CTWS Farmosa Foundation Model

Code Llama

對話問答自動生成 Code

給定一個整列數列，以及任意的值，找出陣列中兩個數合為該值的位址

```
def find_sum_pair(arr, target):  
    for i in range(len(arr)):  
        for j in range(i+1, len(arr)):  
            if arr[i] + arr[j] == target:  
                return (i, j)  
    return None
```

說明：

1. 我們使用兩層循環來遍歷陣列中的每個數字。
2. 在每個循環中，我們都使用 `range` 函數來獲取陣列中從索引 `i` 開始到最後一個元素的範圍。
3. 在每個循環中，我們都使用 `if` 語句來檢查陣列中兩個數字的和是否等於目標值。
4. 如果找到了兩個數字的和等於目標值，我們則返回兩個數字的索引。

Code Llama

IDE 擴充功能：如 Code 填空

Code 自動填空

Debug Code

You

anything wrong? :

```
function minval(arr, n) {  
    let currmin;  
  
    for (let i=0; i<n; i++) {  
        if (arr[i] < currmin)  
            currmin = i;  
    }  
    return currmin;  
}
```

Code Llama

Yes, there is a mistake in the code. The variable **currmin** should be initialized to **Infinity**, not 0. This ensures that the first element of the array will always be considered as the minimum value, even if it's negative.
Here's the corrected version:

```
function minval(arr, n) {  
    let currmin = Infinity;  
  
    for (let i=0; i<n; i++) {  
        if (arr[i] < currmin)  
            currmin = arr[i];  
    }  
    return currmin;  
}
```

Unit Test

You

write unit test :

```
const removeLeadingWhitespace = (text) => {  
    const lines = text ? text.split('\n') : [];  
    let min = Number.MAX_VALUE;  
    lines.forEach(line => {  
        const trimLine = line.trimStart();  
        if (trimLine.length > 0) {  
            const index = line.indexOf(trimLine);  
            if (index < min) {  
                min = index;  
            }  
        }  
    });  
  
    if (min > 0 && min !== Number.MAX_VALUE) {  
        return lines.map((line) => line.replace(new RegExp(`^\\s{${min}}`), ''));  
    }  
    return text;  
};
```

Code Llama

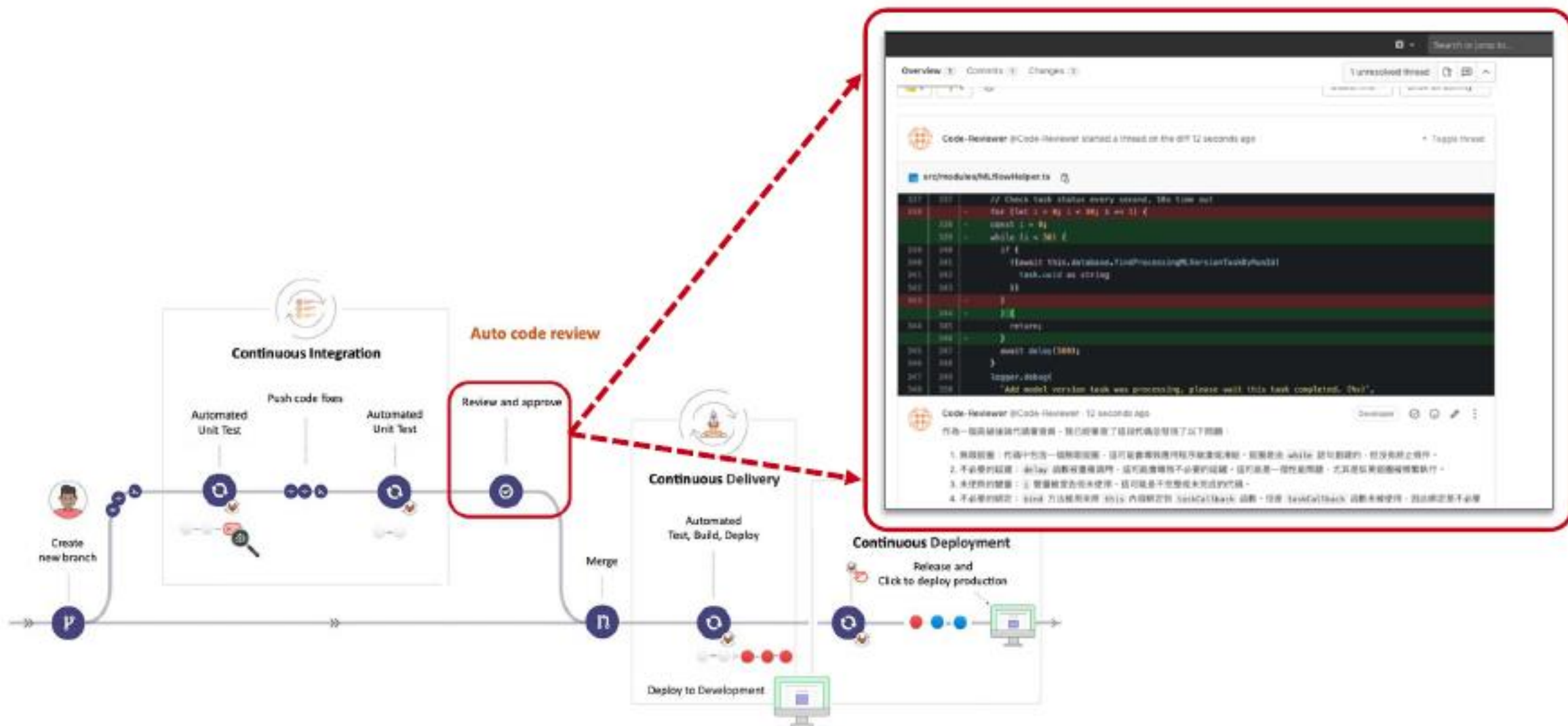
Here is a possible unit test for the `removeLeadingWhitespace` function:

```
describe('removeLeadingWhitespace', () => {  
    it('should remove leading whitespace from each line of a string', () => {  
        expect(removeLeadingWhitespace('  hello\nworld')).toBe('hello\nworld');  
    });  
  
    it('should not modify lines that do not have leading whitespace', () => {  
        expect(removeLeadingWhitespace('hello\nworld')).toBe('hello\nworld');  
    });  
  
    it('should handle empty strings', () => {  
        expect(removeLeadingWhitespace('')).toBe('');  
    });  
});
```

This test suite checks three different cases:

- A string with leading whitespace on some lines, which should be removed by the function.
- A string without any leading whitespace, which should remain unchanged.

AI Code Reviewer on CI / CD pipeline



生成式AI 的發展議題


可信賴AI

- Controlling Process
- Open Source Models
- Edge AI

普惠 AI

- AI for All (Affordable AI)
- Open Innovations
- AI Everywhere

永續 AI

- Green/Low Carbon
 - Monitoring
 - Optimization
- 

開源大型語言模型發展快速: 從完全開源到商用/開源兩個陣營

- Decoder Model
- 訓練方式: 預測下一個字 (文字接龍)
- Unsupervised training
- Pre-trained Model

- 單一模型處理多任務
- In-Context Learning
 - Zero Shot
 - Few Shot
 - Prompt

- Alignment for Users
 - Instruction Tuning
 - RLHF

從此不揭露技術資訊!
Close Source Model

 OpenAI

2018/6
GPT
117M

2019/2
GPT-2
1.5B

2020/5
GPT-3
175B

2022/12
ChatGPT
175B

2023/4
GPT - 4

RNN
LSTM

2017/6
Transformer

2018/10
BERT
250M

Google

2022/7
BLOOM
176B

2023/7
LLAMA 2
70B/2T Data

BLOOM

- 大量平行運算
- Gradient 較不易消失
- 學到的特徵較容易泛化
- Encoder Model
- 訓練方式: 克漏字
- 不同模型處理不同下游任務

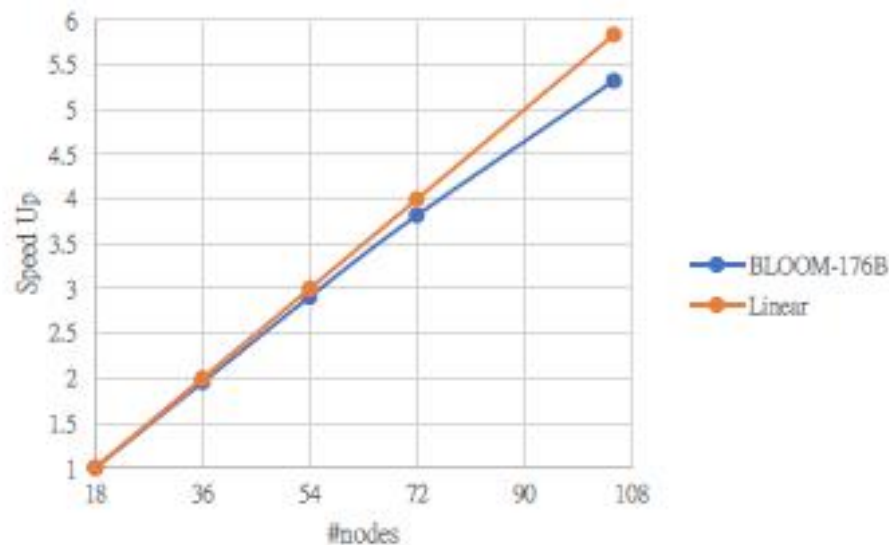
- Decoder model
- Autoregressive
- Open Source



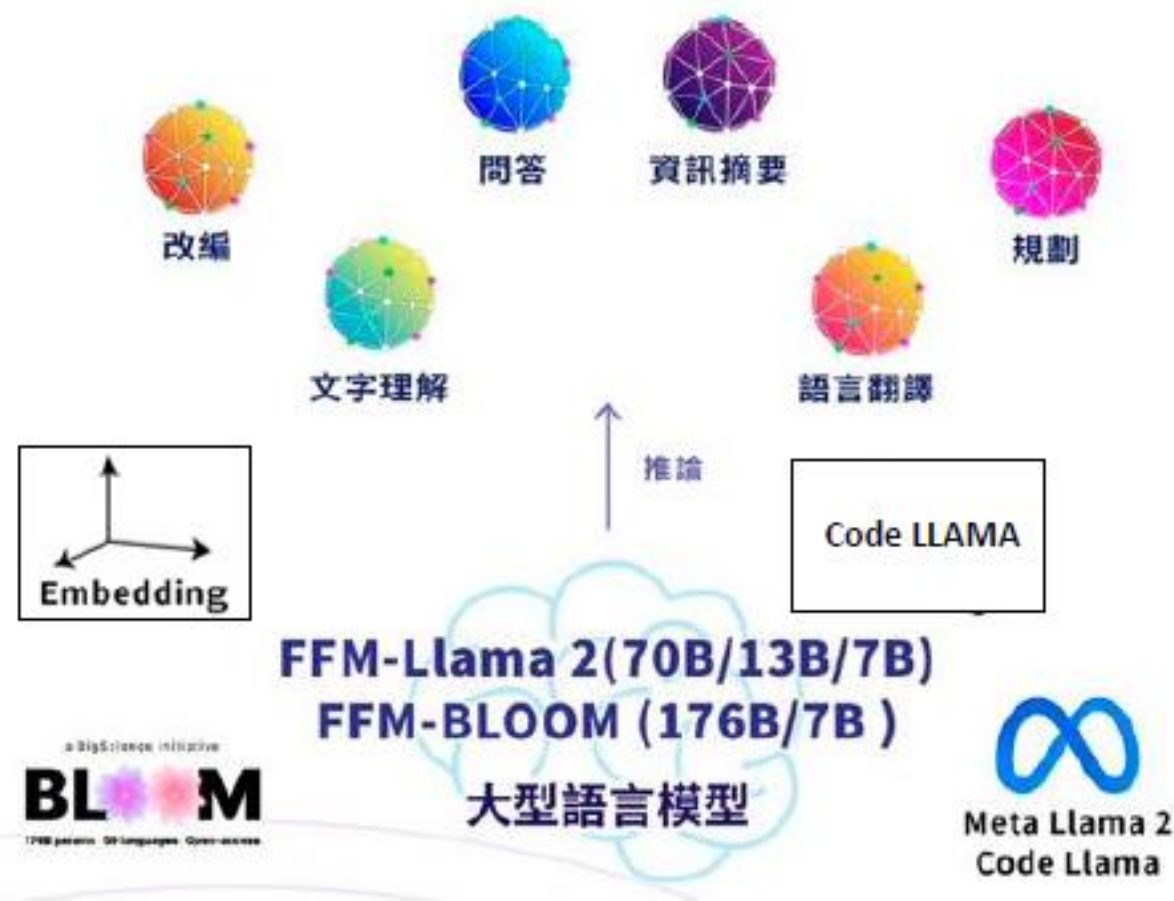
算力即國力!

台灣杉二號AI超級電腦: 台灣發展生成式AI技術的加速器

- 生成式AI 需要以GPU為算力核心的AIHPC
- 全球前100大HPC中共有25座商業目的AIHPC, 台灣杉二號排名64, 算力為9PFLOPS
- 25座商用AIHPC中, 大型石油公司擁有的7座、俄羅斯擁有的5座、NVIDIA 自有的3座, 美國有4座、三星電子一座、SKT一座、歐洲兩座、UAE一座、台灣一座
(真正能讓商用客戶使用到的AIHPC算力稀缺)
- 台灣杉二號的AIHPC as a Service 已可讓客戶以雲服務的體驗使用到AIHPC 資源
- 華碩電腦、台灣大哥大、兆豐金控、永豐金控合作成立台灣智慧雲端服務公司(台智雲)進行商業營運、技術開發與產能持續擴充



發表全球首發繁中福爾摩沙 LLaMA2 全系列多樣化可控大語言模型



- 全球第一個繁體中文強化版的FFM-Llama 2 (70B /13B /7B)全系列模型
- 大幅提升FFM-Llama 2的繁體中文能力，兼具原生Llama 2優異的回應方式和能力
- BLOOM 176B 繁體中文強化模型
- 多樣化開源模型選擇：提供 Llama2、Code Llama、BLOOM、Embedding、等多模型，滿足自然語言處理、文本生成、Code生成各種需求

Japanese



Formosa Foundation Model

PLAYGROUND



日本の政治制度



日本の政治制度は、国会制民主主義を基礎とし、行政府（内閣）、立法府（国会）、司法府（裁判所）の三つの政府が憲法に基づいている。

行政府は、内閣が国家の事務を管理し、内閣総理が政府の首脳である。内閣総理は、国会の議員から選出され、閣僚の助力を受けて政府を運営している。

立法府は、国会が法律を制定し、政府の施政を監視している。国会は、衆議院（下院）と参議院（上院）の二院制で、衆議院の議員は、公選によって選出され、任期は4年である。参議院の議員は、公選によって選出され、任期は6年である。

司法府は、裁判所が法律を解釋し、司法の独立性が保障さ

Vietnamese



Formosa Foundation Model

PLAYGROUND



hệ thống chính trị việt nam



Hệ thống chính trị Việt Nam là hệ thống chính trị do Đảng Cộng sản Việt Nam lãnh đạo, gồm các cơ quan quyền lực nhà nước, các tổ chức chính trị - xã hội và các tổ chức khác. Hệ thống chính trị Việt Nam được tổ chức và hoạt động theo Hiến pháp và pháp luật Việt Nam.

優化開源大型語言模型LLaMa2 70B 的繁中能力

LLaMa2 Open Source

L 生成一個人露營旅行可能需要的十個項目的清單:



Sure, here's a list of 10 items that a person might need for a camping trip:

1. Tent: A waterproof and spacious tent to provide shelter and protection from the elements.
2. Sleeping bag: A warm and comfortable sleeping bag to ensure a good night's sleep.
3. Sleeping pad: A sleeping pad or air mattress to provide extra comfort and insulation.
4. Backpack: A sturdy and comfortable backpack to carry all necessary items.
5. Clothing: Appropriate clothing for the weather conditions, including layers for cold temperatures and rain gear.
6. Footwear: Sturdy and comfortable hiking boots or shoes with good grip for hiking.
7. Headlamp or flashlight: A reliable source of light for

重新生成

福爾摩沙大模型

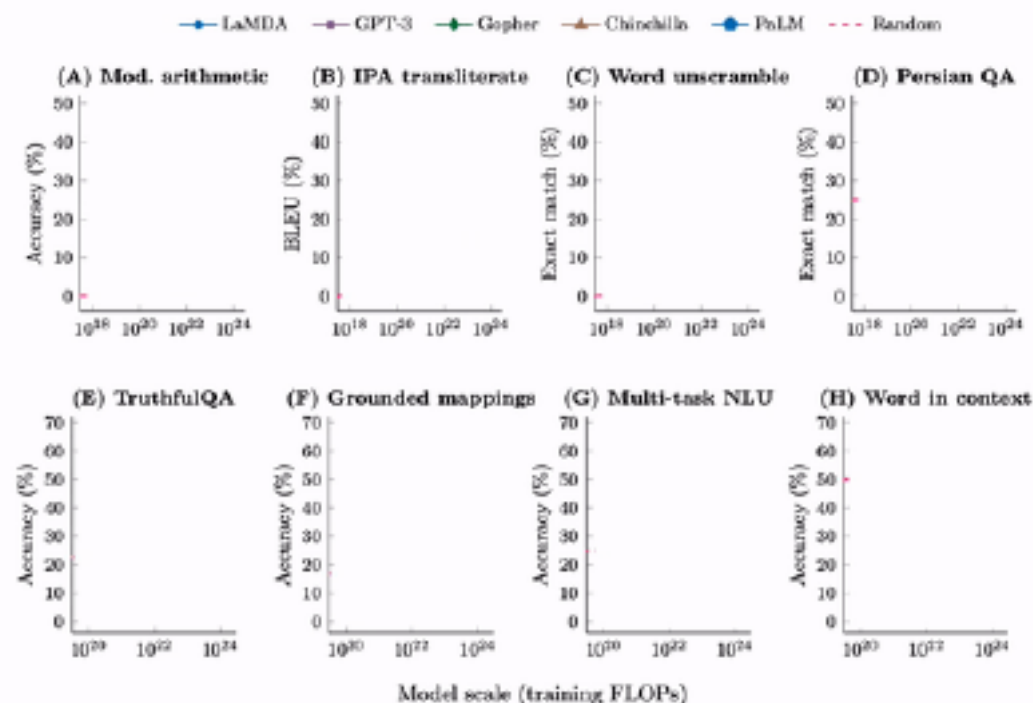
L 生成一個人露營旅行可能需要的十個項目的清單:



1. 帳篷：提供遮雨和遮風的重要運動設備，讓您在戶外住宿時感到舒適和安全。
2. 睡袋：提供舒適的睡眠環境，讓您在戶外住宿時感到舒適和輕鬆。
3. 露營墊：提供遮雨和遮風的重要運動設備，讓您在戶外住宿時感到舒適和安全。
4. 照明設備：讓您在夜晚感到舒適和安全，並且更容易移動和找到您需要的東西。
5. 露營桌：讓您在戶外住宿時有一個方便的地方，放置您需要的東西，並且感到舒適和方便。
6. 露營椅：提供舒適的坐墊，讓您在戶外住宿時感到舒適和輕鬆。
7. 廚房設備：讓您在戶外住宿時能夠煮食烹茶，有一個方便的廚房環境。
8. 水壺和儲水袋：讓您在戶外住宿時有足夠的水，並且方便攜帶和儲存。
9. 防蚊帳：讓您在戶外住宿時能夠擺脫蚊子的威脅，感到舒適和安全。
10. 露營裝備包：讓您能夠輕鬆攜帶和儲存您所需的所有露營設備和用品，方便您的旅行和住宿。

大模型具有模型湧現的能力

大模型的跑分及理解能力皆較佳



Model	Size	Code	Commonsense Reasoning	World Knowledge	Reading Comprehension	Math	MMLU	BBH	AGI Eval
LLAMA 2	7B	16.8	63.9	48.9	61.3	14.6	45.3	32.6	29.3
	13B	24.5	66.9	55.4	65.8	28.7	54.8	39.4	39.1
	34B	27.8	69.9	58.7	68.0	24.2	62.6	44.1	43.4
	70B	37.5	71.9	63.6	69.4	35.2	68.9	51.2	54.2

福爾摩沙大模型 70B

北風與太陽 北風與太陽兩方為誰的能量大相互爭論不休。他們決定，誰能使得行人脫下衣服，誰就勝利了。北風一開始就猛烈地刮，路上的行人緊緊裹住自我的衣服，風見此，颯得更猛。行人冷得發抖，便添加更多衣服。風颯疲倦了，便讓位給太陽。太陽最初把溫和的陽光灑向行人，行人脫掉了添加的衣服，太陽之後把強烈陽光射向大地，行人們開始汗流浹背，漸漸地忍受不了，脫光了衣服，跳到了旁邊的河裡去洗澡。

請解析上述故事中想呈現給讀者的寓意。

故事中寓意的內涵是，溫柔的方式常常比用力更有效。在故事中，北風試圖用強大的刺激去讓行人脫衣服，但是行人反應卻把更多的衣服增加在身上。然而，太陽則是用溫暖和愉快的陽光，最終讓行人感到舒適，並脫光衣服。這個故事告訴我們，用輕柔的方式去達成目標，通常比用力或是施壓更有效。

透過雲端服務訓練與優化模型、無須投入巨額資本

- ✓ 容易上手的 **No-code Portal**
- ✓ 以小時計價、便捷調用**AIHPC** 超級電腦算力
- ✓ 已支援FFM-BLOOM-7B與FFM-BLOOM-176B
- ✓ 9月底上線FFM-LLaMA2 7B/13B/70B
- ✓ 9月底上線FFM-BLOOM 支援Bitfit
- ✓ 10月上線FFM-LLaMA2 、Meta LLaMA2 與FFM-BLOOM系列, 支援LoRA PEFT
- ✓ 也會陸續上線更多夥伴開發的基礎模型與模型壓縮功能



實際案例：以數十萬筆 RMA 維修資料進行大語言模型微調，來洞察與預測產品的使用期限與維修成本

基於 FFM-Llama-7B 經過 Fine-tuning 所誕生的『智慧製造預防性保修模型』

透過產品的“模組類別”、“組件類別”和“銷售日期”，來判斷該產品售出之後過多久會需要維修

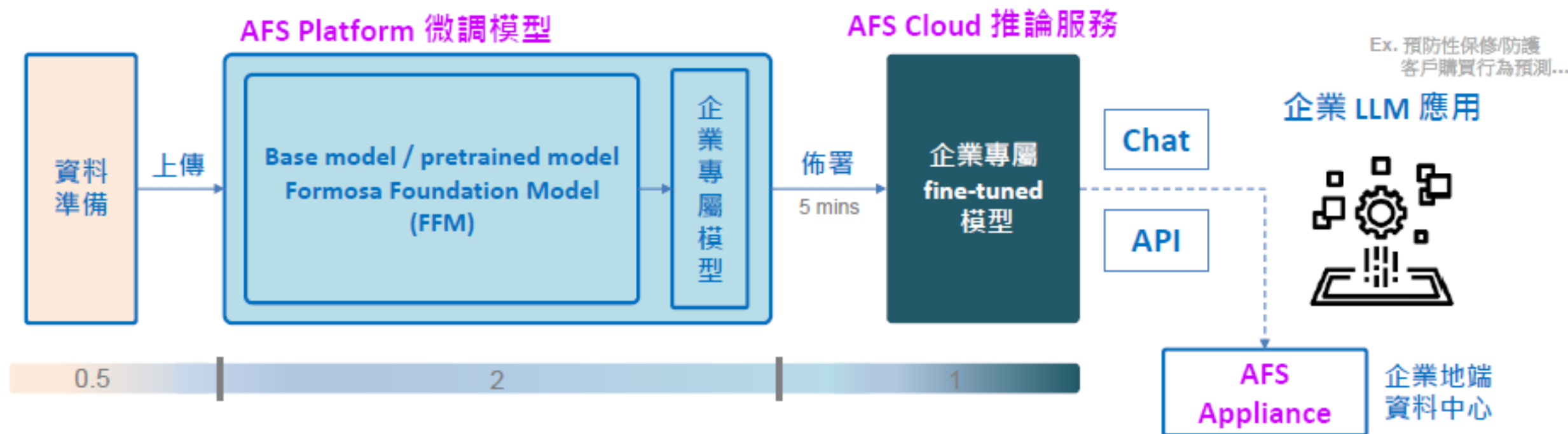
模型微調過程 僅需數小時，且實際產品使用時間與模型預測產品使用時間平均誤差約 2.7 個月

這個模型 帶來的效益 可以用來評估給予產品的保固條件，以及作為評估維修品所需的庫存量

而在 未來 相同性質的預測任務可以透過此流程打造出 企業專屬的x任務預測模型

Ex. 客戶的購買模式預測, 生產設備的預防性維護...

企業微調專屬模型之流程



此次模型從資料準備到驗證完畢約 3.5 hours (不包含資料收集)

實際微調時程依據不同的資料內容與大小有所改變

華碩產品歷史維修資料，該檔案中每個Row都是一個產品的維修紀錄，包括產品模組類別(M)、組件類別(P)、銷售的月份/年份、該產品維修的月份/年份、維修的組件數量。資料區間為2005年2月至2009年12月。

	A	B	C	D	E
1	module category	component category	year/month(sale)	year/month(repair)	number repair
2	M6	P16	2007/9/1	2009/4/1	1
3	M2	P30	2007/9/1	2009/8/1	1
4	M1	P12	2006/10/1	2008/2/1	2
5	M1	P30	2006/5/1	2007/7/1	1
6	M3	P06	2007/8/1	2007/12/1	1
7	M7	P19	2006/7/1	2007/6/1	1
8	M7	P04	2006/3/1	2008/4/1	1
9	M3	P09	2006/11/1	2007/6/1	1
10	M2	P22	2007/3/1	2009/9/1	1
11	M2	P28	2007/8/1	2008/5/1	1
12	M5	P12	2006/1/1	2006/5/1	1
13	M3	P02	2006/10/1	2007/9/1	1
14	M2	P06	2007/10/1	2008/5/1	1

```
[{"role": "system", "content": "You are a data analyst. Each product will have module_category and component_category and sales date. You will judge the life cycle of this product based on module_category, component_category and sales date."},
```

```
{"role": "human", "content": "The module_category is M1, component_category is P02, sales date is 2006/4/1. Please analyze the life cycle of product."},
```

全部內容為一筆資料

```
{"role": "assistant", "content": "The life cycle of product is 15 months after sales date."}]
```

System : 模型風格設定

Human : 希望模型預測的輸入資料

Assistant : 希望模型回答的答案

總資料筆數 : 約 數十萬筆

訓練資料筆數 : 90%

訓練資料Token數 : >50M

測試資料筆數 : 10%

```
1 [{"role": "system", "content": "You are a data analyst. Each product will have module_category and component_category and sales date. You will judge the life cycle of this product based on module_category, component_category and sales date."}, {"role": "human", "content": "The module_category is M1, component_category is P02, sales date is 2006/4/1. Please analyze the life cycle of product."}, {"role": "assistant", "content": "The life cycle of product is 15 months after sales date."}]
2 [{"role": "system", "content": "You are a data analyst. Each product will have module_category and component_category and sales date. You will judge the life cycle of this product based on module_category, component_category and sales date."}, {"role": "human", "content": "The module_category is M1, component_category is P02, sales date is 2006/6/1. Please analyze the life cycle of product."}, {"role": "assistant", "content": "The life cycle of product is 13 months after sales date."}]
3 [{"role": "system", "content": "You are a data analyst. Each product will have module_category and component_category and sales date. You will judge the life cycle of this product based on module_category, component_category and sales date."}, {"role": "human", "content": "The module_category is M1, component_category is P02, sales date is 2006/5/1. Please analyze the life cycle of product."}, {"role": "assistant", "content": "The life cycle of product is 12 months after sales date."}]
4 [{"role": "system", "content": "You are a data analyst. Each product will have module_category and component_category and sales date. You will judge the life cycle of this product based on module_category, component_category and sales date."}, {"role": "human", "content": "The module_category is M1, component_category is P02, sales date is 2006/7/1. Please analyze the life cycle of product."}, {"role": "assistant", "content": "The life cycle of product is 10 months after sales date."}]
5 [{"role": "system", "content": "You are a data analyst. Each product will have module_category and component_category and sales date. You will judge the life cycle of this product based on module_category, component_category and sales date."}, {"role": "human", "content": "The module_category is M1, component_category is P02, sales date is 2006/10/1. Please analyze the life cycle of product."}, {"role": "assistant", "content": "The life cycle of product is 14 months after sales date."}]
6 [{"role": "system", "content": "You are a data analyst. Each product will have module_category and component_category and sales date. You will judge the life cycle of this product based on module_category, component_category and sales date."}, {"role": "human", "content": "The module_category is M1, component_category is P02, sales date is 2006/9/1. Please analyze the life cycle of product."}, {"role": "assistant", "content": "The life cycle of product is 11 months after sales date."}]
7 [{"role": "system", "content": "You are a data analyst. Each product will have module_category and component_category and sales date. You will judge the life cycle of this product based on module_category, component_category and sales date."}, {"role": "human", "content": "The module_category is M1, component_category is P02, sales date is 2006/9/1. Please analyze the life cycle of product."}, {"role": "assistant", "content": "The life cycle of product is 12 months after sales date."}]
8 [{"role": "system", "content": "You are a data analyst. Each product will have module_category and component_category and sales date. You will judge the life cycle of this product based on module_category, component_category and sales date."}, {"role": "human", "content": "The module_category is M1, component_category is P02, sales date is 2006/4/1. Please analyze the life cycle of product."}, {"role": "assistant", "content": "The life cycle of product is 14 months after sales date."}]
9 [{"role": "system", "content": "You are a data analyst. Each product will have module_category and component_category and sales date. You will judge the life cycle of this product based on module_category, component_category and sales date."}, {"role": "human", "content": "The module_category is M1, component_category is P02, sales date is 2006/5/1. Please analyze the life cycle of product."}, {"role": "assistant", "content": "The life cycle of product is 13 months after sales date."}]
```

模型微調 – 輕鬆無負擔，滑鼠在手，專屬模型快速擁有

透過AI Foundry Service (AFS) Platform，您只需要使用滑鼠，將資料上傳後，建立起微調任務，接著等待訓練完成，您就可以獲得您的專屬模型，整體流程完全以No Code方式達成。

TWCC

TAIWAN
COMPUTING
CLOUD

AFS-T01-AF... ▾

服務列表 ▾

⚙️ 📞 ⓘ

🌐 AFS 服務

資料集

Platform

Shuttle

Cloud

主頁 > AFS > Platform > 詳細資料

Platform 任務詳細資料

配置

🗑️ 刪除

🔄 重新整理

基本資訊

ID	fb8110ac-d4ed-4fa9-aa70-bcad3d6aa600	建立時間	2023-11-02 15:39:55
名稱	rmallama2dc2	結束時間	2023-11-02 17:45:12
描述	建立者		
狀態	● completed		

推論驗證 – 不只學習到了資料間的關聯性，賦予給模型的任務需求使命必達

A The module_category is M9, component_category is P31, sales date is 2006/8/1. Please analyze the life cycle of product.

 The life cycle of product is 19 months after sales date.

依照設定訓練的內容(格式)進行答覆

```
[{"role": "system", "content": "You are a data analyst. Each product will have module_category and component_category and sales date. You will judge the life cycle of this product based on module_category, component_category and sales date."}, {"role": "human", "content": "The module_category is M1, component_category is P02, sales date is 2006/4/1. Please"}, {"role": "assistant", "content": "The life cycle of product is 15 months after sales date."}]
```

訓練資料範本

關閉工具欄

新對話

發送訊息...

預設指令

You are a data analyst. Each product will have module_category and component_category and sales date. You will judge the life cycle of this product based on module_category, component_category and sales date.

確認

不只訓練、推論也需要算力

模型訓練階段

CNN/RNN (< million parameters)

AI 1.0
決策式AI



模型推論階段

CPU/GPU/NPU



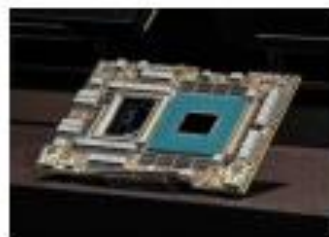
1000X
~1000000X
算力要求

Transformer-based LLM (> billion parameters)

AI 2.0
生成式AI



H100



GH200



AIHPC

GPU/NPU (Memory is key)

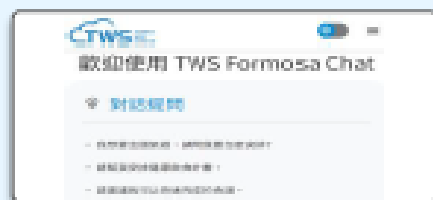


L40S

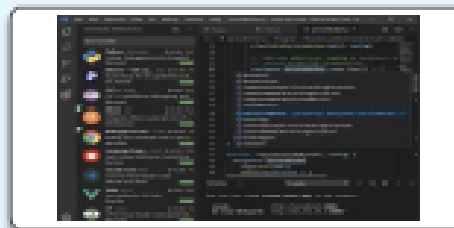


更多樣貌的LLM 推論算力on PC/Server: 與台灣優勢PC/Server供應鏈共創成長

多樣化應用及服務



AI 助手



程式 助手



RAG 應用

統一 API 介面

TWSC LLM Universal API

多元模型

LLM 模型 (Llama 2 / Code Llama/ BLOOM / Embedding)

推論引擎

TGI

OpenVINO

llama.cpp

加速硬體

GPU

iGPU/NPU

CPU

Server / Workstation

Intel MTL

一般 PC / Notebook

多樣態終端



AI Server



AI Workstation



ZenBook 14



ExpertBook



MacBook



NUC

台灣目前已可利用雲端服務快速佈署開源大型語言模型

- 用戶只需支付GPU 算力費用, 便可快速啟動多種基礎模型
- 方便企業客戶驗證基礎模型能力 (可結合台智雲免費額度), 也可以讓企業客戶長期使用。
- 將陸續上線更多Open Source Model 與夥伴提供的基礎模型



讓企業自主在各環境佈署開源大型語言模型

資安

機敏資料

合規

隱私

透過 AFS Pipeline 完成的企業專屬模型可選擇 AFS Appliance 服務，將模型佈署於地端資料中心



ON-PREMISE

企業資料中心應用系統

無縫整合

Inference
API



AFS Appliance



ESCN8



ESC8000A



ESC4000A

華碩 ESC-N8、ESC8000系列、ESC4000系列 AI 伺服器提供最佳化的算力

生成式AI 驅動Knowledge App 典範轉移



Bill Gates
微軟創辦人

1980年來最重要技術進步！
與GUI一樣會改變世界
It's a shock, this thing is amazing



Geoffrey Hinton
AI 教父
圖靈獎得主

AGI(通用型人工智慧)
藉由 LLM 達成



Satya Nadella
微軟 CEO

This is a game changer

- Next generation AI Infrastructure
- 對最大應用 Search 的影響
- Office, Teams, Copilot



Sergey Brin
Google 創辦人

穀歌創始人正式回歸，加入
LLM戰爭！



陸奇
奇績創壇創辦人兼CEO
微軟, 百度, 雅虎經理人

接下來15-20年，模型就是知識，將無處不在。以後手機上打開，任何聯網，模型就過來了。它教你怎麼去解答法律問題，怎麼去做醫學檢驗，不管什麼樣的模型都可以無處不在。

我們當下的位置

- 模型與人和社會的關係
- 拐點後新範式的發展格局

第二代系統

模型無所不在

第一代系統

資訊無所不在

通用智能

湧現、代理
功能可見、具身

共同進化

模型（知識）
生態系統

資訊（感知）
生態系統

