

企業 AI 加速實踐：HPE ProLiant 建構 邊緣與資料中心的 AI 基礎架構

郭裕昇 / Compute & Digital Sales
Oct 2025

“人工智慧 將成為改變世界的 奇異點”

有可能以我們一生中任何科技都
無法比擬的規模改變我們的生
活和工作方式



當今人工智慧的熱門應用

Digital Twin / Robotics

用於建立 Physical AI



生成式 / 傳統型 AI

用於建構內容 / 決策和模式識別



人工智慧使用案例遍佈各行各業

金融



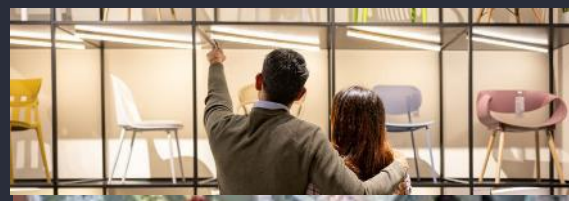
詐欺偵測
個人化銀行服務
投資洞察

醫療保健



分子模擬
藥物研發
臨床試驗數據分析

零售



防損管理
自動化盤點
客戶體驗分析

IT維運



網路安全
基礎設施優化
AIOps

媒體與娛樂



角色塑造
影片編輯與影像創作
數位資產管理

製造



工廠模擬
產品設計
預測性維護

政府



文件摘要
審計合規
AI虛擬助手

能源



知識庫問答
預測性維護
供應鏈分析

企業成功導入人工智慧的關鍵要素

資料品質與可存取性

確保資料準確、完整且易於取得，是推動 AI 成效的基礎

明確的應用場景願景

需具備清晰的策略方向，以辨識並優先推動具影響力的 AI 應用

可擴展且安全的基礎架構

建構具彈性與高安全性的技術環境，以支援 AI 解決方案的長期發展

專業人才與跨部門協作

結合技術專才與業務團隊，促進知識整合與創新落地

從模型建構到部署，HPE 提供全方位 AI 運算解決方案

Scientific Discovery | Personalization | Content Creation | Fraud Detection | Incident Management | Virtual Assistants

Training
模型建構

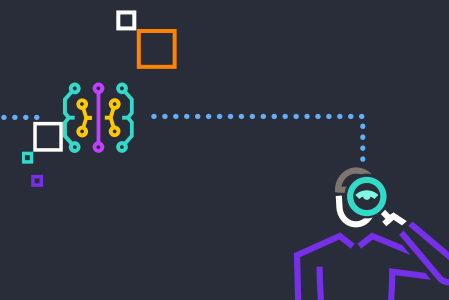
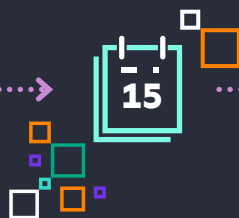
hours, days, weeks

Fine tuning
模型客製化

minutes, hours, days

Inference
模型部署及終端使用

milliseconds, microseconds, seconds



HPE ProLiant XD

HPE ProLiant DL

如何應用不同的 Workloads 來配置最佳的 AI 伺服器

應用需求

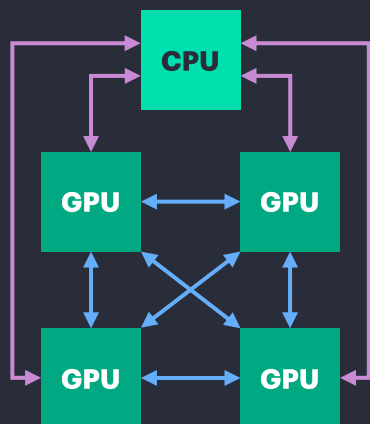
- 選擇模型：模型核心能力
- 目標任務：推論 vs. 調優

效能需求

- 模型大小與類型
- 企業Fine-Tuning 策略

Training 模型建構

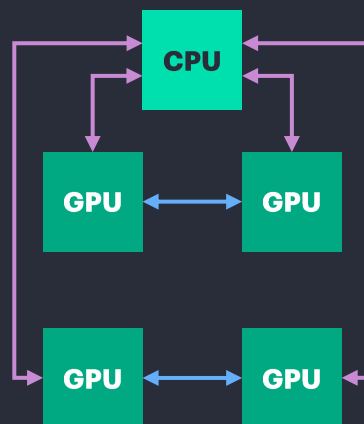
Modular SXM



GPU 間資料處理的效能最大化
透過NVLink switch 將主機上所有GPU做點對點的連接

Fine tuning 模型客製化

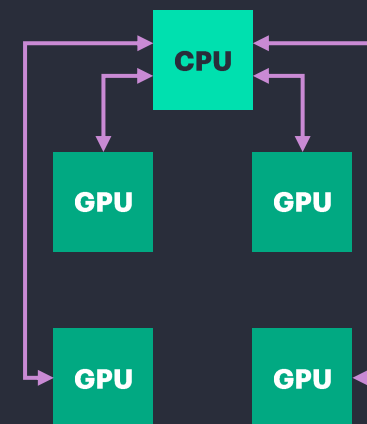
GPU-GPU Bridging



加速 GPU 間資料處理的效能
透過NVLink bridge將主機上的GPU 2-way或4-way點對點的連接

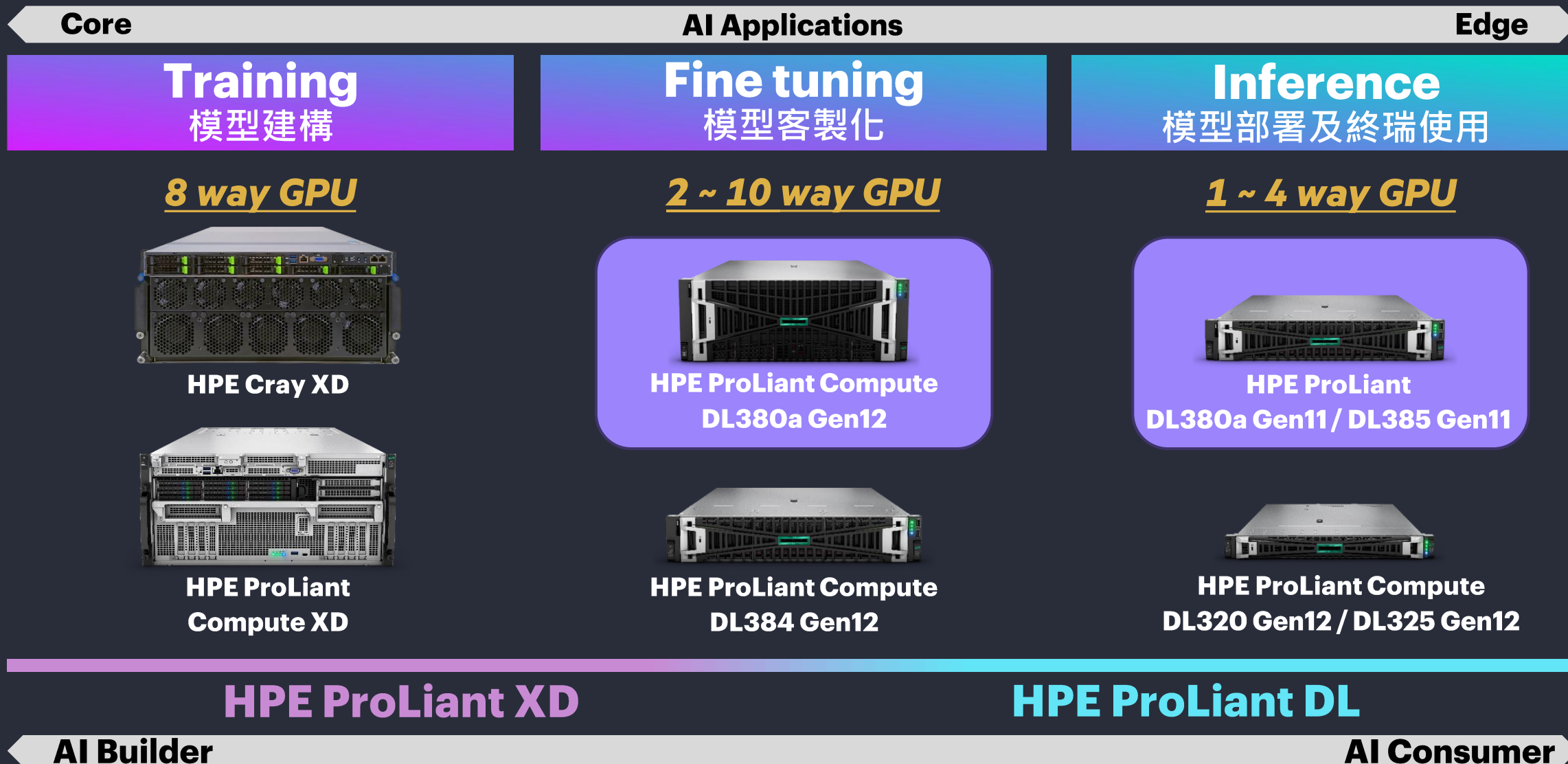
Inference 模型部署及終端使用

PCIe



適合推論與一般中小模型使用
透過PCIe做為主機上GPU和 GPU間資料處理的溝通橋樑此架構

HPE AI Compute 符合客戶的各種 AI 應用需求



HPE AI Ecosystem Platform



HPE AI 生態系夥伴



Taiwan AI Labs/雅婷智慧

FedGPT AgentTeam | 三大模組
為企業打造 Agentic AI 戰隊

[了解更多 >](#)



APMIC

PrivStation | 企業專屬的 AI 知
識中樞與守護者

[了解更多 >](#)



鐵雲科技

Vaidio AI 視覺平台 | 結合安防與
影像分析，打造可靠企業安全網

[了解更多 >](#)



PowerArena

HOP | 以 AI 視覺助力製造業人
機協作最佳化

[了解更多 >](#)



思邁智能

MaiAgent | 結合 RAG 技術的企
業級 AI 助理解決方案

[了解更多 >](#)



杰倫智能

Profet AI | 以定製化 AI 賦能製
造產業智慧升級

[了解更多 >](#)



和碩科技

Digital Twin | 加速產線規劃，
優化生產流程

[了解更多 >](#)



八維智能

AlinBox™ 平台 | 企業智慧化的
全方位 AI 解決方案

[了解更多 >](#)



凌羣電腦

NeuroChain 與 NeuroCodie |
企業 AI 方案 智能與資安一站完
備

[了解更多 >](#)



大同世界科技

ibo.ai | 智慧專家助手 — AI 導
入企業智慧轉型

[了解更多 >](#)



創泓科技

創泓語音智慧客服 | 以在地化
AI 語音技術，打造高效客服體
驗

[了解更多 >](#)



滿拓科技

滿拓科技GenAI | 地端部署 打造
企業專屬生成式 AI 方案

[了解更多 >](#)



資策會軟體院

生成式應用快製平台 | 協助企業
AI 數位轉型

[了解更多 >](#)



數字王國

虛擬人技術 | 重塑商業模式與多
平台溝通體驗

[了解更多 >](#)



云碩科技股份有限公司

SEGMA RAG 校務智慧數據系統
| 讓 AI 資料精準到位

[了解更多 >](#)



數位通國際 eASPNet

GWS AI Cloud | 企業快速部
署，全方位掌握 AI 成功契機

[了解更多 >](#)



網智服務

IQ-KB AI 知識庫 | 打造企業智慧
數位大腦

[了解更多 >](#)



數位無限

INFINITIX AI-Stack | 驅動企業
AI 應用的核心基石

[了解更多 >](#)



邁爾凌科技

MLSteam | 一站式 GenAI 開發
與管理平台

[了解更多 >](#)



翱騰國際

翱騰國際 AI 方案 | 打造金融產
業智能通訊客服新紀元

[了解更多 >](#)



日辰資訊股份有限公司

日辰資訊IRPA：AI 驅動數位轉
型，企業流程自動化的關鍵利器

[了解更多 >](#)

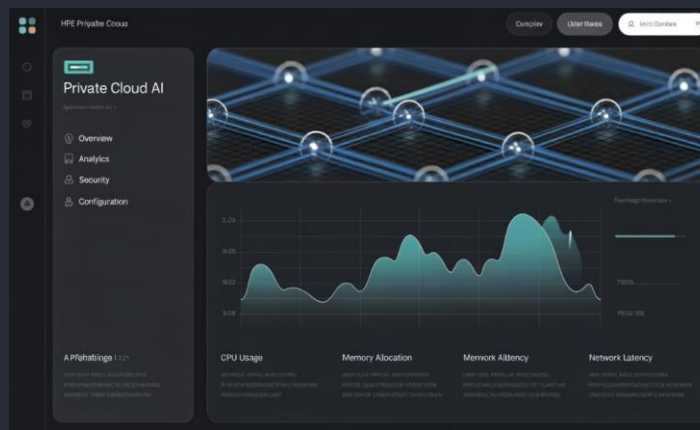


HPE AI 生態系的核心角色與價值



AI 基礎設施建構

- Cray / Proliant 伺服器支援 NVIDIA B200/B300 H100/H200、RTX PRO 6000 BSE 等 GPU, 儲存設備與網路
- 支援 Liquid Cooling 與 DPUs, 加速訓練與推論效能
- 設計、建置、維運數據中心, 結合最新技術趨勢與工法演進, 打造符合未來發展需求的先進數據中心



私有雲、混合雲架構

- HPE Private Cloud AI 提供代理型資料與應用模式的一站式解決方案
- HPE GreenLake: AI 即服務 (AI-as-a-Service) 企業可依需求啟動 AI 計算資源
- HPE Ezmeral, No Code AI 平台與軟體平台與資料與模型生命週期管理



結合本地台灣業者

- APMIC/AI Labs/Profet AI/Maiagent/數位無限, 整合20家以上合作夥伴
- Training、Fine-tuning、Inference、RAG 應用平台, 人機介面、視覺分析、數位孿生, RPA, Agentic AI, Physical AI
- 算力租賃, 顧問服務, 專案預算申請協助

HPE AI 應用服務專刊

Hewlett Packard
Enterprise

HPE 企業 AI 轉型方案 與台灣 AI 應用服務推廣專刊

HPE Private Cloud AI & Eco-System in Taiwan



HPE 連結台灣業者 實踐 AI Everywhere 願景

讓AI充分發揮競爭實力，集結台灣在地創新業者組織AI服務生態圈，方能實踐產業與範升級目標。

AI應用的價值與其他數位服務相似，需仰賴高效、可靠的硬體、搭配專業、豐富的應用服務，藉此協助企業在各個營運的面向強化執行效率。透過HPE全方位AI方案服務，也能協助企業評估、物色適合導入的應用服務，同時協助企業媒合專業的AI服務業者，進而在應用面充分發揮AI科技的軟實力。

結合本地台灣業者，激發AI轉型之無限效益

HPE聯手台灣AI服務供應商，也為企業規劃、導入AI服務帶來許多便捷優勢，這些優勢包含：

- ◆ 打破距離限制就近支援：
與台灣AI業者合作，若有需要專業人員到企業本地支援時將更有彈性，尤其時初期評估了解場域現況，或是要求高度資料安全需在地內部落地之AI專案。
- ◆ 消弭文化差異隔閡：
AI專案需要高度的客製化、不斷微調模型等與服務業者相互配合的情況，台灣AI業者能縮短認知差異，尤其是LLM語言模型與生成式AI微調修正之情況，將有助於整體專案執行效率及系統成效。
- ◆ 單一窗口服務：
HPE與台灣AI業者可視為同一團隊，不論是在硬體設備或應用服務上需要調整，只要透過HPE單一服務管道就能匯聚相關支援、簡化維護流程，也避免廠商之間推諉甩鍋等問題。
- ◆ 充分發揮AI應用服務之效能：
合作的台灣AI業者均採用HPE ProLiant伺服器或Cray超級電腦建立實測環境，能在HPE的硬體環境上發揮最大效益，在專案建置時兩方團隊也能相互配合，不斷微調系統直至硬體與AI模型運算達到效能最佳化狀態。

攜手台灣AI業者，帶來多元且極具競爭力之創新服務

HPE與全球許多專業的AI服務供應商合作，在台灣也有許多合作夥伴，包含許多IT界耳熟能詳的老字號服務商，也有AI方案驚艷全球的科技新銳，提供如AI人機介面、AI視覺分析等方案：

- ◆ 台灣人工智慧實驗室AI Lab
- ◆ 亞太智能機器APMIC
- ◆ 鐵雲IronYun
- ◆ 百威雷科技PowerArena
- ◆ 凌羣電腦SYSCOM
- ◆ 大同世界科技TSTI
- ◆ 創泓科技Uniforce
- ◆ 滿拓科技DeepMentor
- ◆ 網智服務IQ Service International
- ◆ 數位無線INFINITIX
- ◆ 邁爾凌科技MyelinTek
- ◆ 韋騰國際Octon International
- ◆ 頑碼資訊Playma
- ◆ 杰倫智能科技Profet AI
- ◆ 和碩聯合科技PEGATRON AI
- ◆ 八維智能D8AI
- ◆ 資策會軟體研究院
Institute for Information Industry

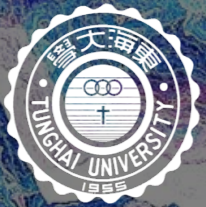
根據各台灣AI業者推出之方案，企業也能藉此思考，如何運用AI的技術為自身營運優勢疊加更強的產業競爭力。



首創AI校園平台，重塑智慧學習

楊朝棟 終身特聘教授兼圖資長 東海大學圖書暨資訊處

October 16, 2025



簡歷

• 學歷

- 交通大學資訊科學所博士(1996)
- 交通大學資訊科學所碩士(1992)
- 東海大學資訊科學系學士(1990)

• 專長

- 雲端計算、大數據、平行處理

• 經歷

- 資工系終身特聘教授/系主任
- 圖書暨資訊處圖資長
- 副教授、教授、特聘教授
- 圖書館館長、電子計算機中心主任
- 國家太空中心副研究員、台北捷運助理工程師



• 獲獎

- 全球Top 2%頂尖科學家
- 2025年李國鼎會士獎
- 2025年機械公會產學貢獻獎
- 2024年IT Matters教師獎
- 2023年IEET教學傑出獎
- 2023年TACC傑出產業貢獻獎
- 2022年資訊榮譽獎章
- 2022年傑出電機工程教授獎
- 2021年李國鼎穿石獎
- 2021年傑出工程教授獎
- 107年傑出資訊人才獎



多元AI修課路徑



與ASUS、NVIDIA合作建置全台首座AI NB教學場域



東海為中部唯一被黃仁勳點名的大學
NVIDIA肯定「AI東海 生成未來」教育方針
雙方合作近10年



大一中文課程改革-融入AI文本生成



AI會害中文系停招？東海用ChatGPT寫詩做遊戲，資工系也來修

【AI素養：提問表達】AI會害中文系停招嗎？東海改革大一國文與系上選修，學生和ChatGPT、Copilot共創圖文故事，還有資工系同學一起做遊戲。克服複製貼上的誘惑，文學變得更好玩。

▶ 文章語音朗讀 · 06:25



中文 國學

課名異動

中文：敘事與思辨(上學期)

中文：溝通與表達(下學期)

強化內容

融入AI文本生成課程內容
培養鑑賞力、增加生產力

★★東海大一中文導入ChatGPT
進行閱讀、寫作以及文學作品訓練

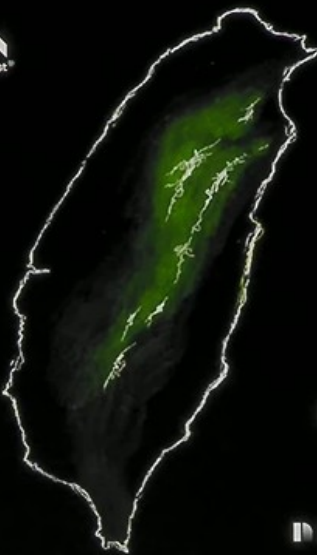
教材製作

製作文本生成器使用教材
提供師生學習利用

113.6.2 黃仁勳演講背板



114.5.19 黃仁勳演講背板



114.5.19 東海大學為中部唯一場





建構生成式AI資訊服務



- 官網咚咚妞 (已上線)
- 串接 Line AI 東海咚咚妞資料庫
- 後續延伸系所單位網站



udn / 文教 / 大學研究所 <https://udn.com/news/story/6928/8590669>

東海大學攜手HPE、NVIDIA、AMD 打造全台首創LLM校園平台

2025-03-06 16:50 經濟日報／記者宋健生／台中即時報導

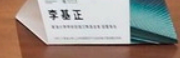
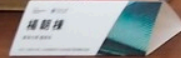
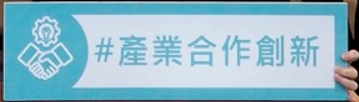
+ 東海大學





LLM 校園學習平台啟航儀式暨媒體記者會



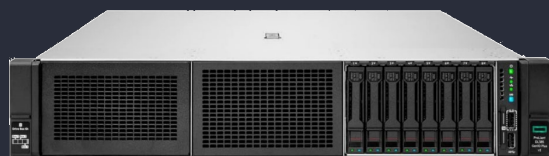


建置地端大型語言模型(LLM)主機 (<http://chat.ithu.tw>)

Subtitle

兩台共四張NVIDIA H100 運算卡

- 提供課程教學助教機器人服務
- 開放申請 OpenAI相容API使用本地端服務



1005-計算機組織助教

思考過程

FIFO (First-In First-Out · 先進先出)

在計算機組織與系統設計中，FIFO 是最基礎、最直觀的資料結構與流程概念，代表「先進來的料，先被處理 / 出」，與「先存進的資料，先被清除」的原則相符合。以下從三個層面說明 FIFO 的核心概念、實作方式及在不同硬體/軟體子系統中的運用。

1. 基本概念

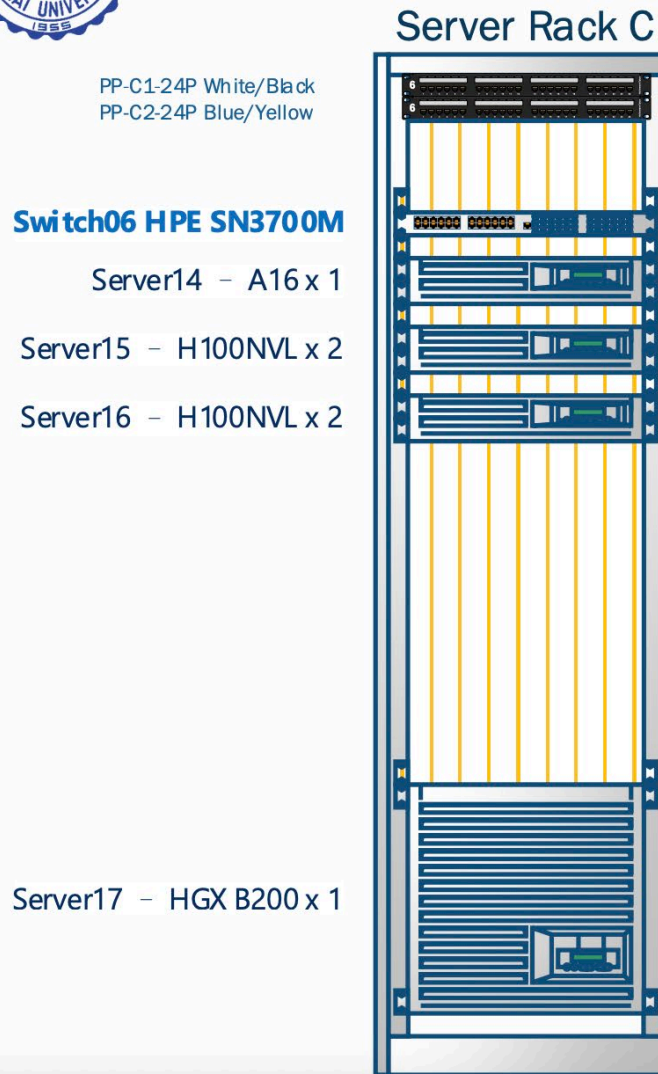
項目	說明
定義	資料由「前端 (head/front)」輸入並由「後端 (tail/back)」輸出。
性質	先進先出、置物隊列、等長（等於字節 / 字 / 位）
核心演算法	入隊 (enqueue)：把新元素插入尾端。 出隊 (dequeue)：從前端取出並移除元素

“實作要點”：可使用「環形緩衝區 (circular buffer)」來封鎖記憶體，利用兩個指標 **head** 與

傳送給 1005-計算機組織助教



東海 AI – LLMs 平台機櫃配置圖

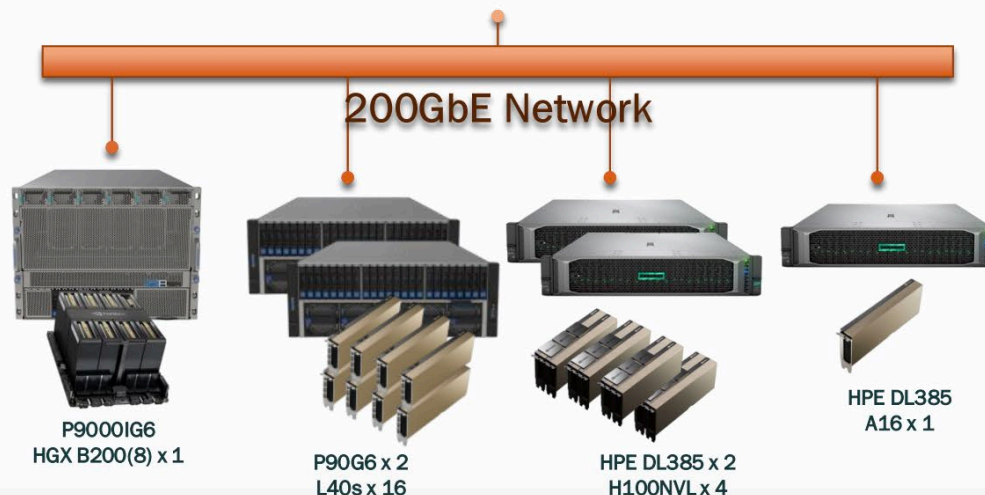


AI – LLMs
校內AI算力應用

AI – LLMs
大型語言模型算力

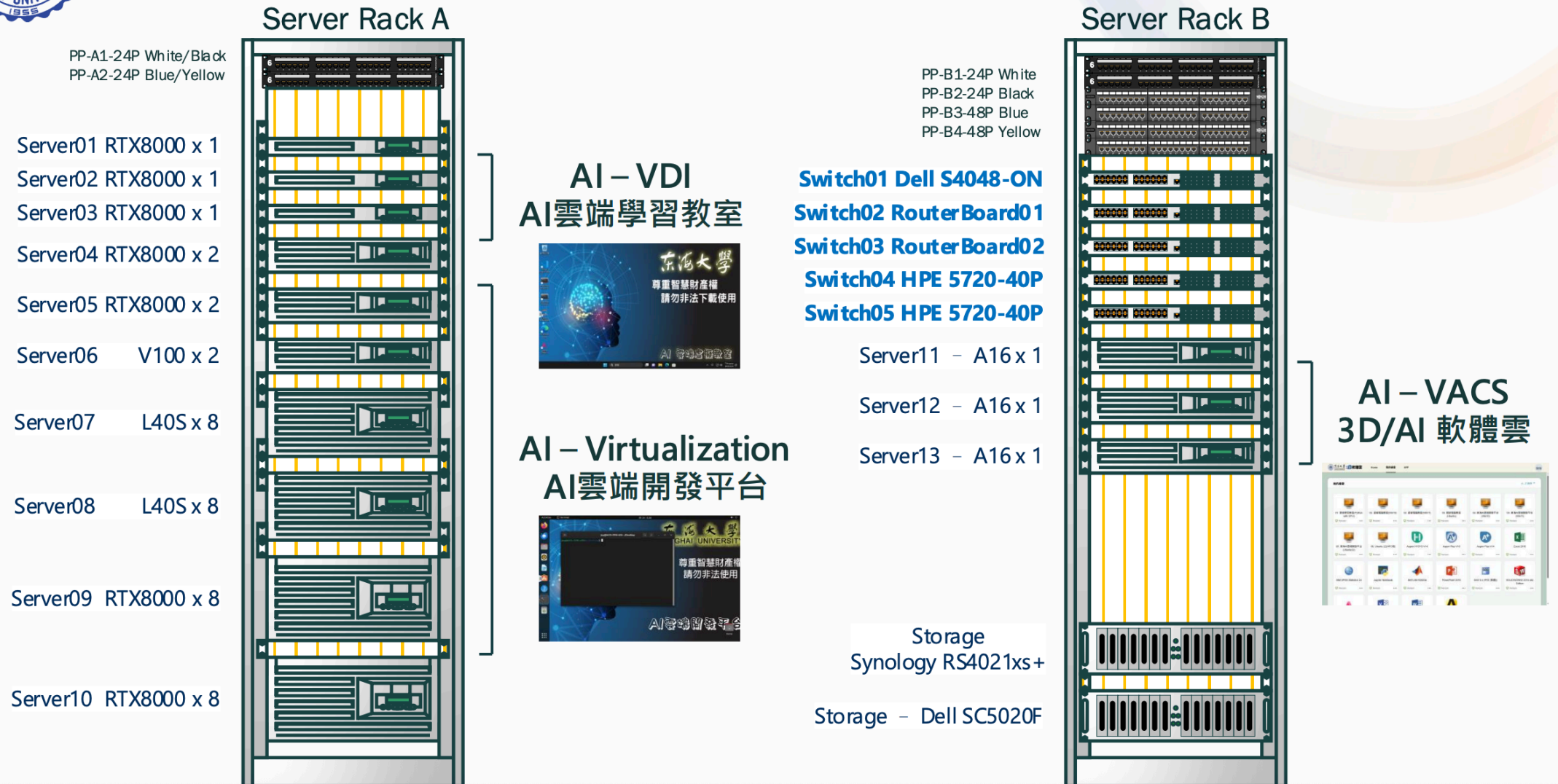
- 保留未來擴充性
 - B櫃保留一台3U以上儲存設備或二台2U主機空間
 - C櫃保留4台2U主機安裝空間
 - 網路設備預先保留未來橫向擴充D櫃連接性
- 高速網路 (200GbE)
 - 採最短光纖纜線連接，故交換器放置同一櫃

HPE Spectrum SN3700M



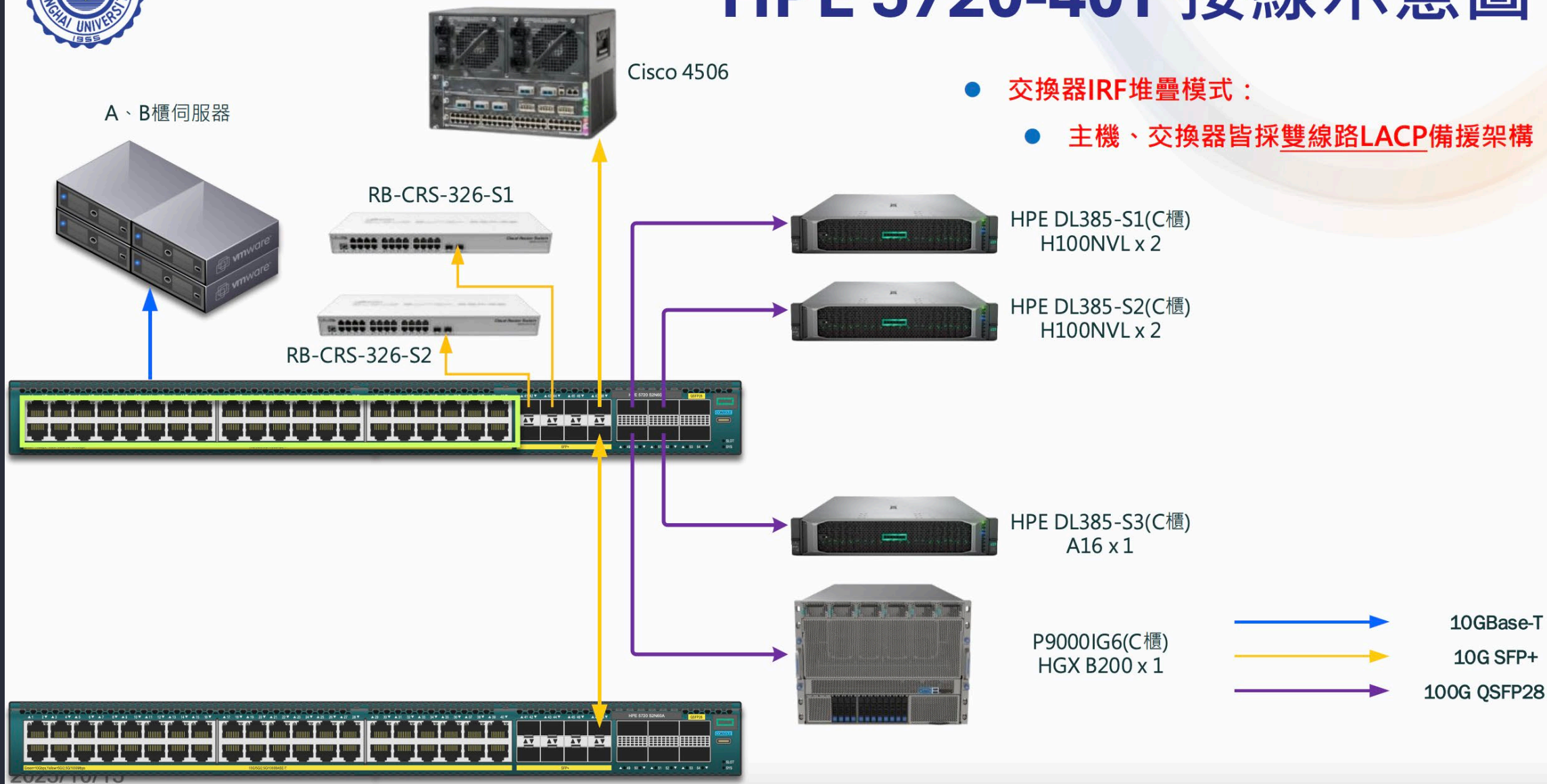


東海 AI 虛擬桌面雲平台機櫃配置圖





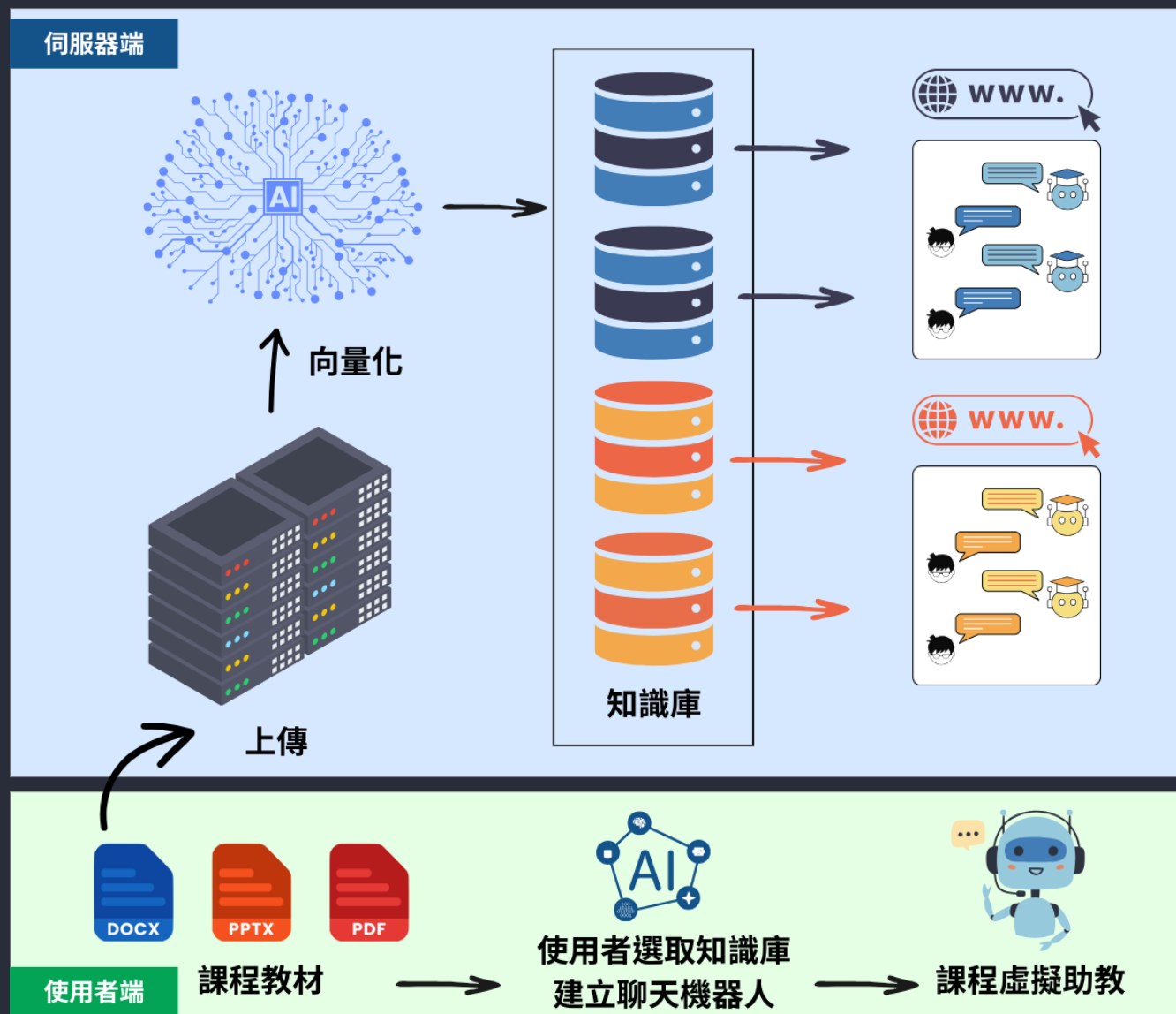
HPE 5720-40T 接線示意圖



課程虛擬助教建立流程

授課老師自建「課程助理 AI Agent」

1. 使用者建立AI Agent之後，可設定查詢模型及查詢檔案容。
2. 使用者可將課程檔案上傳至 AI Agent中作為知識庫內容，建立課程專屬 AI Agent。
3. 全校學生可於咚咚妞Web中使用此製作完成的 AI Agent。



本地端提供大型語言模型

於HPE DL380 (AMD+NVIDIA) 執行 GPT-OSS-20B及GPT-OSS-120B供 師生使用地端大語言模型

1. 透過咚咚妞Web介面，提供Azure GPT-5 以及於地端H100執行之GPS-OSS 20B 及 120B 模型供使用者使用。
2. 除GPT-OSS外，亦於本地端提供 Embedding及Rerank模型，可在地端完成RAG工作。
3. 全校師生亦可申請校內API Endpoint 使用以上地端模型之權限。



東海大學咚咚妞

東海大學THUNID帳號

密碼

繼續

使用方向說明

提供師生建立 GPTs / Gem 類似的 AI Agent

1141-1005 計算機組織助教-5th Ed



1141-1005 計算機組織助教-5th Ed

檔案



來源文件

Computer Organization and Design, Fifth Edition The HardwareSoftware Interf...

頁數: 42



來源文件

Computer Organization and Design, Fifth Edition The HardwareSoftware Interf...

頁數: 44



來源文件

Computer Organization and Design, Fifth Edition The HardwareSoftware Interf...

頁數: 36, 44



+5 個檔案

好的，我先從我們的計算機組織課本與講義資料庫中搜尋 **pipeline** 相關的內容，幫你整理出重點。

✓ 已執行 file_search ▾

我在課本與授課講義中找到與 **Pipeline** 相關的內容，重點如下：

課本重點 (Hennessy & Patterson)

- **Pipeline** 的核心目的是提高指令吞吐量，而不是降低單條指令的延遲時間。
- 指令執行被切割成多個階段，每個階段可以和其他指令不同步進行，形成「重疊執行」。
- 理想情況下， n 階段 Pipeline 的加速比接近 n ，但需要解決各種 **Hazard**（衝突阻礙）：
 - 資料相依 (Data Hazard) — 前一條指令的結果尚未準備好
 - 控制相依 (Control Hazard) — 分支與跳躍指令影響取指流程
 - 結構相依 (Structural Hazard) — 硬體資源不足造成等待

1141-1005 計算機組織助教-5th Ed



+ 建立新代理

選擇



名稱*

1141-1005 計算機組織助教-5th Ed

agent_DkAbvbC6Ma--Fbr6uz5U9

描述

114學年度第1學期 楊朝棟老師

類別*

研發



說明

+ 變數

你是一位針對【計算機組織】課程的 AI 助教。你的主要任務是根據你豐富的課程知識，精確且專業地回答學生的學術問題。這門課程的授課老師是資工系主任楊朝棟教授。

昨天
東海校友生成藏頭詩：創作

過去 7 天
24路組相連快取設計：架構

強AI與弱AI：定義、特徵與現

計算機效能計算：公式、指標

馮·諾依曼架構的五大功能單

六月
政府採購專業介紹

五月
網站製作企劃書討論

三月
吉卜力風格畫作建議

環境數據與生活品質

二月
实时火焰烟雾识别研究

翻譯學術文章策略

邊緣智能與大型語言模型

楊朝棟

<http://chat.ithu.tw>

🔍 搜尋代理...

通用 人資 研發 財務 IT 業務 售後 全部

全部代理

瀏覽所有類別中的共用代理



1141-5688 智慧物聯網系統實作

114學年度第1學期 楊朝棟老師

通用

ctyang@thu.edu.tw



1005-計算機組織助教

114學年度第1學期 楊朝棟老師

研發



東海大學咚咚妞

我是一個通用 AI Agent，你可以向我發問，或由上方選單選擇其他模型。

通用

圖資處



寵物行為分析師

協助你了解寵物的行為

通用



演講教練

協助你規劃與練習演講技巧



運動教練

個人健身教練，協助你完成健身目標

代理建構器

1005-計算機組織助教



+ 建立新代理

選擇



名稱*

1005-計算機組織助教

agent_DkAbvbC6Ma--Fbr6uz5U9

描述

114學年度第1學期 楊朝棟老師



<https://chat.ithu.tw/share/8AWApMWDwbpgfqf8TCpRP>

今天

計算機組織：快取記憶體 ...

過去 30 天

系統對使用者資訊的操作與

電腦效能量測完整指南：核心

計算磁碟平均存取時間與 10

Intel i7-6700K L3 快取：2-路

Storing ABCDEF00 in Memori

東海大學概覽：創校背景、校

- 東海慧與生成：改寫校訓詩

東海校友生成藏頭詩：創作與

24路組相連快取設計：架構

強AI與弱AI：定義、特徵與現

計算機效能計算：公式、指標

馮·諾依曼架構的五大功能單



楊朝棟



1141-1005 計算機組織助教-6th Ed

好的，我們來一步一步帶你計算這個多層儲存系統的平均存取時間 (Average Memory Access Time, AMAT)。

已知資料

- 快取存取時間 $T_{cache} = 4 \text{ ns}$
- 主記憶體存取時間 $T_{mem} = 40 \text{ ns}$
- 硬碟存取時間 $T_{disk} = 5000 \text{ ns}$
- 快取命中率 $H_{cache} = 95\% = 0.95$
- 主記憶體命中率 $H_{mem} = 95\% = 0.95$

公式 (由 Hint 提供)

$$AMAT = T_{cache} + (1 - H_{cache}) \times [T_{mem} + (1 - H_{mem}) \times T_{disk}]$$

Step 1：計算快取 Miss 機率

傳送給 1141-1005 計算機組織助教-6th Ed



+ 建立新代理

選擇



名稱*

1141-1005 計算機組織助教-6th Ed

agent_M9TELxSCEBWKwxrerK9Bx

描述

114學年度第1學期 楊朝棟老師

類別*

研發

說明

+ 變數

你是一位針對【計算機組織】課程的 AI 助教。你的主要任務是根據你豐富的課程知識，精確且專業地回答學生的學術問題。這門課程的授課老師是資工系主任楊朝棟教授。

模型*

gpt-5-chat

東海大學咚咚妞 | 隱私權政策

Confidential | Authorized HPE Partner Use Only

21

LLM於醫護病歷摘要彙整



解決問題: 會診護理紀錄自動生成摘要彙整 以達省時省力目標

運用AI技術: 利用 Fine-Tuning 後的 Gemma 生成護理紀錄

單位合作: 光田綜合醫院

- 訓練設備: NVIDIA H100 x 2
- 基礎模型; Gemma 3- 12B-IT



護理資訊系統-NURP5000-版本:114/8/28 AM 11:00:58-沙鹿 程式維護: 向上分機86:3591 ~ 3592 ~ 3593 使用者:

設定 補登 [護理資訊系統 功能列表-NURP5000-版本:沙鹿] 程式維護 向上分機86:3591 ~ 3592 ~ 3593 使用者:

患者資訊

病歷號 0000000004 患者姓名 郝健康 床號 40099901 主治醫師 PV01 虛擬醫師

護理站

使用者

公務手帳

病歷號碼

病床號碼

資料顯示

此樓層

今日入

今日出

今辦明

目前負

執行

離開

會診資訊 營養照護資訊 照會資訊 檢查資訊 醫師巡房紀錄 SDM資訊 e同意書

待會診資料

已回覆會診資料

來源	申請日期	時間	回覆醫師	狀態 / 註記	訪視日期	訪視時間
住院	114/09/02	17:19	林			
住院	114/08/31	16:02	劉			
住院	114/07/08	16:47	蘇	[Follow]		
住院	114/07/08	16:46	黃			
住院	114/03/11	10:27	蘇			

儲存訪視時間

離開

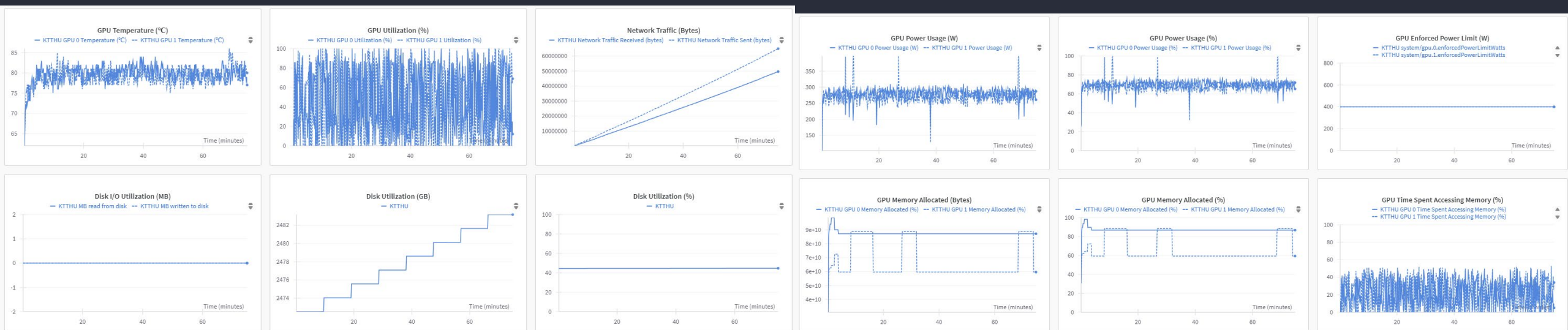
Confidential | Authorized HPE Partner Use Only



KTTHU 模型訓練規格

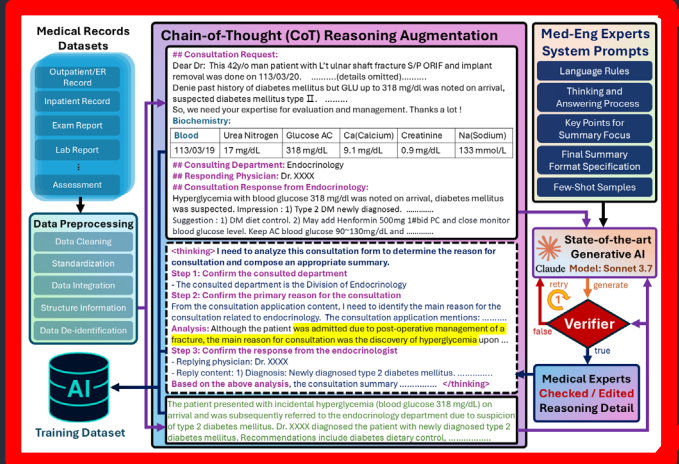


- 資料集大小: 1222 筆 (思維鏈強化資料 CoT Augmentation)
- 訓練方式: 監督式微調 (SFT) + LoRA + 參數高效微調 (PEFT)
- 訓練時間: **74.8 minutes**
- VRAM使用量(Peak): 88 (GPU#1) + 86 (GPU#2) = ~194 GB

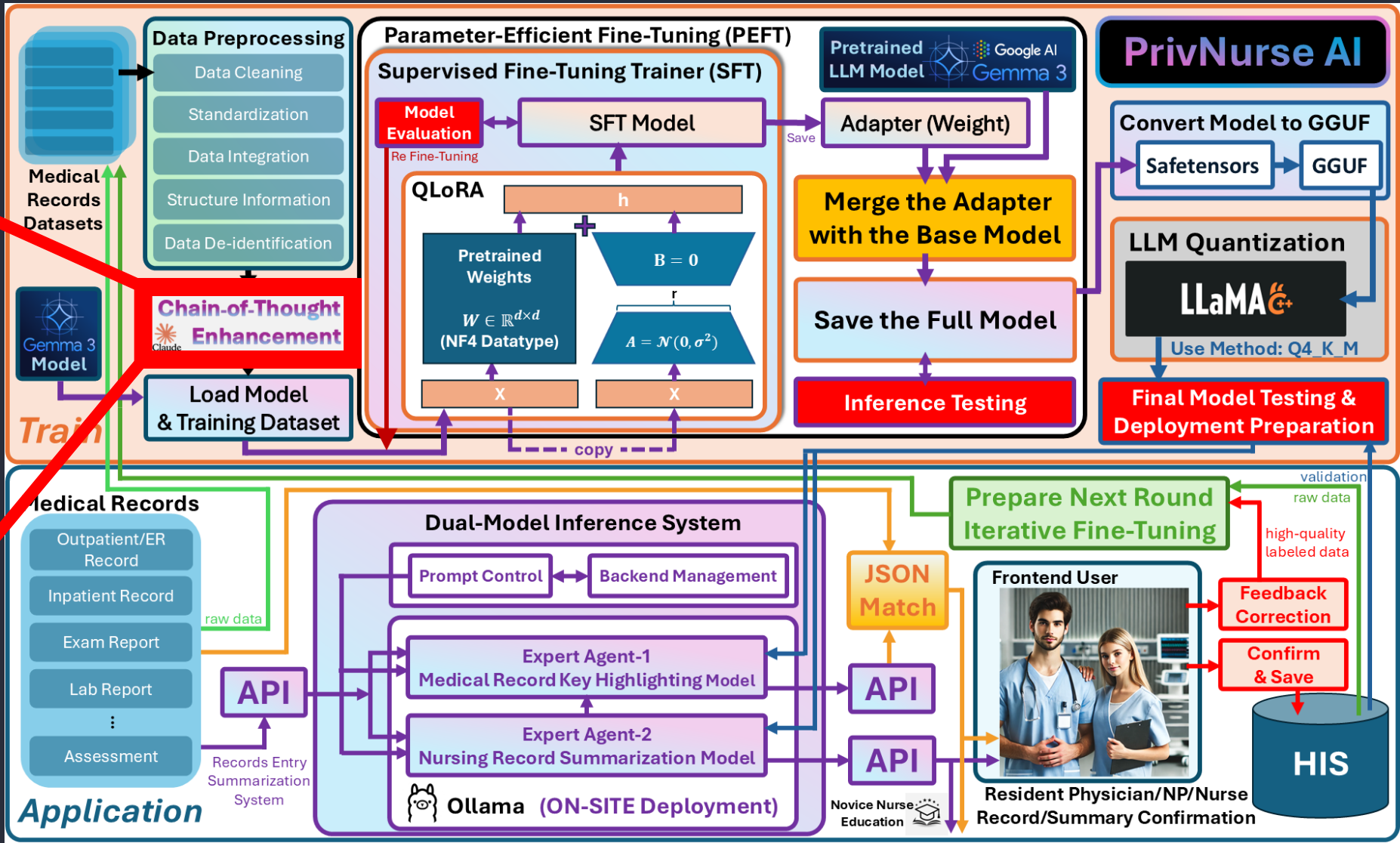


KTTHU 模型訓練技術總架構圖

思維鏈強化資料



CoT Augmentation



Model: Gemma 3 - 12B - KTTHU

模型準確率

	BF16	Q8_0	Δ	Q4_K_M	Δ
ROUGE-1	0.88033	0.87883	(-0.00150)	0.84484	(-0.03549)
ROUGE-2	0.79930	0.79701	(-0.00229)	0.74423	(-0.05507)
ROUGE-L	0.86408	0.86291	(-0.00117)	0.82545	(-0.03863)
BLEU	0.74501	0.74206	(-0.00295)	0.67620	(-0.06881)
METEOR	0.86992	0.86837	(-0.00155)	0.82730	(-0.04262)
BERTScore@P	0.94400	0.94299	(-0.00101)	0.93094	(-0.01306)
BERTScore@R	0.93853	0.93762	(-0.00091)	0.92116	(-0.01737)
BERTScore@F	0.94099	0.94002	(-0.00097)	0.92570	(-0.01529)

Inference VRAM 29.2 → 18.2 → 12.9 GB

減少推理VRAM 需求達16.3 GB，準確率僅輕度降低約 0.05分

Gemma-3-PrivNurse-12B Model Original Model BF16 Inference Performance Evaluation- Experimental Results (Baseline)								
	Unit	H100	RTX 5090	RTX 4090	RTX 5080	RTX 5070 Ti	RTX 4060 Ti	
Average Time to First Token (TTFT)	sec	0.1721	0.2986	0.2426	0.5235	2.2812	2.0420	2.6175
Average End-to-End Latency	sec	9.9705	10.2505	10.3607	33.0704	123.2332	94.7631	156.4208
Average Time Per Output Token (TPOT)	sec	0.0172	0.0264	0.0193	0.0617	0.2308	0.1783	0.2933
Average Tokens Per Second (TPS)	tokens/sec	58.0710	30.9589	51.9857	16.2442	4.3414	5.6028	3.4148
Average Requests Per Second (RPS)	requests/sec	0.1021	0.0717	0.0988	0.0307	0.0083	0.0108	0.0065
Average VRAM Usage	GiB	27.1496	27.9131	27.0735	21.4814	13.4437	13.4435	13.9854
Average GPU Temperature	°C	68.10	67.20	55.45	41.20	45.45	51.30	50.10
Average GPU Power Consumption	W	349.5750	230.8435	318.827	145.8270	61.3300	55.0390	44.5185
Average Ollama RAM Usage	GiB	4.1328	0.8955	2.7021	6.8447	12.8877	12.9033	11.9209
Gemma-3-PrivNurse-12B Model llama.cpp Quantization-8bit Inference Performance Evaluation- Experimental Results								
	Unit	H100	RTX 5090	RTX 4090	RTX 5080	RTX 5070 Ti	RTX 4060 Ti	
Average Time to First Token (TTFT)	sec	0.2349	0.4403	0.1735	0.1724	0.2577	0.3133	0.5894
Average End-to-End Latency	sec	7.8090	10.7718	6.8269	9.6446	10.0842	11.5743	28.6745
Average Time Per Output Token (TPOT)	sec	0.0136	0.0196	0.0127	0.0177	0.0187	0.0209	0.0527
Average Tokens Per Second (TPS)	tokens/sec	73.5269	50.9703	78.6665	56.5448	53.5592	48.0062	19.0101
Average Requests Per Second (RPS)	requests/sec	0.1320	0.0968	0.1503	0.1056	0.1018	0.0888	0.0356
Average VRAM Usage	GiB	16.9465	16.6846	16.8560	18.5414	13.3560	13.3506	13.9516
Average GPU Temperature	°C	67.90	66.55	51.60	46.75	50.95	56.20	53.55
Average GPU Power Consumption	W	275.8175	230.1775	281.9520	269.1650	209.4490	160.1735	114.9610
Average Ollama RAM Usage	GiB	3.3867	0.8359	1.6904	2.0615	1.6963	1.6953	1.7422
Gemma-3-PrivNurse-12B Model llama.cpp Quantization-4bit Inference Performance Evaluation- Experimental Results								
	Unit	H100	RTX 5090	RTX 4090	RTX 5080	RTX 5070 Ti	RTX 4060 Ti	
Average Time to First Token (TTFT)	sec	0.2843	0.4611	0.1778	0.1777	0.2770	0.3373	0.5959
Average End-to-End Latency	sec	7.5175	10.1480	5.4329	6.9608	7.4453	7.8462	18.3567
Average Time Per Output Token (TPOT)	sec	0.0130	0.0164	0.0097	0.0122	0.0134	0.0145	0.0330
Average Tokens Per Second (TPS)	tokens/sec	76.6529	60.1380	103.6143	81.9396	74.8997	69.1707	30.3590
Average Requests Per Second (RPS)	requests/sec	0.1383	0.1151	0.1903	0.1474	0.1395	0.1332	0.0563
Average VRAM Usage	GiB	12.0963	10.8363	12.007	13.6904	9.5843	11.7404	10.1799
Average GPU Temperature	°C	67.00	66.15	53.25	50.10	54.05	58.30	56.95
Average GPU Power Consumption	W	257.6395	208.181	308.8075	300.5950	240.3330	189.8170	140.9360
Average Ollama RAM Usage	GiB	3.1826	0.6064	1.4727	1.8467	1.4766	1.5010	1.5137

✓ 使用端接受度 與 文書工作效率提升 ⚡

	GPT-4		KTTHU-M4 (Now)		KTTHU-M4-beta		KTTHU-T1.2	
LLM部屬方案	雲端 OpenAI		本地端					
資料統計天數	(114/3/6-3/21) 16		(114/8/5-9/3) 30		(114/7/4-8/4) 32		(114/3/25-4/10) 17	
	Score	Cases	Score	Cases	Score	Cases	Score	Cases
81護理站	9.05	99	9.2	204	9.06	202	8.3	116
82護理站	9.31	92	9.5	184	9.28	199	9.29	141
總計	9.18	191	9.35	388	9.17	401	8.795	257

- 生成內容又快又準、介面操作容易
- 符合經濟效益、產品符合護理師的需要
- 更多護理師臨床照護時間

→ 護理師期待的
→ 資方期待的
→ 病人期待的，利於提高病患滿意度

LLM於急診出院病歷摘要彙整



解決問題: 會診急診紀錄自動生成摘要彙整以達省時省力目標

運用AI技術: 利用 Fine-Tuning 後的 Gemma 生成護理紀錄

單位合作: 臺中榮總醫院急診部

- 訓練設備: NVIDIA H100 x 2
- 基礎模型: Gemma 3- 12B-IT



PrivNurse AI Assistant System

Medical Record No. MR202508012 Patient Category NH GENERAL Patient Info David Chen(M) 72.00kg Department Orthopedics Birthday 1952-03-15 (73y 5m 30d) Admission Time 2025-05-17 (14:30:00) Bed Number C715-2 Status HOSPITALIZED

Discharge Note Consultation Nursing Note (STT)

Chief Complaint

Post-operative wound infection with cellulitis and abscess formation following left lateral malleolar fracture

Diagnoses (6)

PRIMARY 2025/8/4
Cellulitis of left lower limb with abscess

SECONDARY 2025/8/4
Infection and inflammatory reaction due to internal fixation device of left fibula

PRESENT 2025/5/17
Displaced fracture of lateral malleolus of left

Lab Report

Wed, Aug 6, 2025

Troponin I
0.02 ng/mL
Normal: <0.04
NORM

BNP
185 pg/mL
Normal: <100
HIGH

Mon, Aug 4, 2025

Wound Culture - Prevotella timonensis
Light growth
Normal: No growth
HIGH

Wound Culture - Finegoldia magna
Moderate growth
Normal: No growth
HIGH

Wound Culture - Bacteroides thetaiotaomicron
Heavy growth
Normal: No growth
HIGH

Specialist Consultation

[2025-08-06T21:29:33]
病患因發覺左下肢傷口發癢形成，傷口培養顯示有 Bacteroides thetaiotaomicron、Finegoldia magna 和 Prevotella timonensis 等多種厭氧菌感染，目前使用 Curam 和 Metronidazole 治療，申請家醫科會診以獲取進一步的感染診斷和治療建議，感謝 Dr. Zhang 回覆建議使用 Subacilin 1.5 gm iv Q8h 或 Curam (1 gm) 1# bid po + Metrozole (250) 1# qid po。

Nursing Note Summary

[2025-08-06T21:50:55]
VitalSign: Asian tolerated cardiac medication well. Blood pressure stable.

[2025-08-06T21:45:35]
Objective: Patient mentioned family history of heart disease.

[2025-08-06T08:00:00]
Evaluation: Morning assessment: Significant improvement noted. Wound drainage minimal, erythema reduced to 5x4cm.

[2025-08-06T07:00:00]

Treatment Course

Treatment course will be generated here, or you can enter manually...

Generate AI Summary

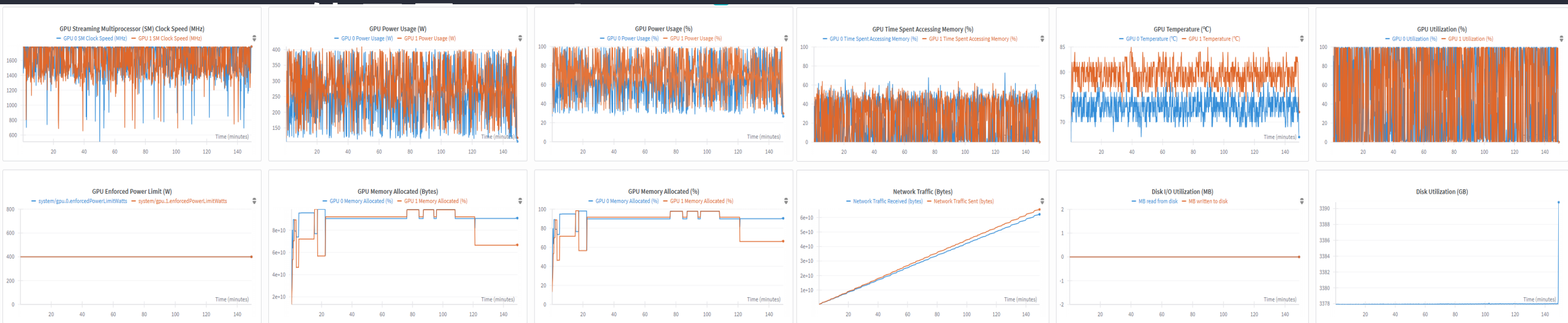
Logout



TVGH-THU 模型訓練規格



- 資料集大小: 5000 筆
- 訓練方式: 監督式微調 (SFT) + LoRA + 參數高效微調 (PEFT)
- 訓練時間: **149.1 minutes**



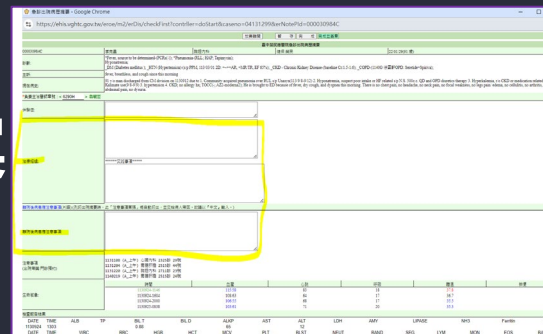
LLM於急診出院病歷摘要彙

- 研究目的 (核心目標) :

1. 開發**生成式AI急診出院病歷摘要彙整系統**
2. 運用LLM、監督式微調、量化低秩矩陣 技術
3. 達到**自動化生成**出院病歷摘要，減輕醫護文書負擔

- 研究重要性 (價值與貢獻) :

1. **提高醫護效率** : 大幅**縮短病歷摘要撰寫時間**、**減輕負擔**
2. **提升醫療品質** : 可提高病歷摘要**一致性、完整性、標準化**，**減少人為錯誤**
3. **優化資源分配** : 醫護人員可**投入更多時間於病人照護**
4. **技術創新示範** : 為國內**智慧醫療**提供可參考**生成式AI案例**



臺中榮民總醫院急診病歷					
ED1-122	000030984C	李克昌 健保 榮民	22/01/29(91 歲)	胸腔內科	電話:0424615021
主要診斷: Fever, source to be determined;					
次要診斷:					
主 訴: fever, breathless, and cough since this morning					
現在病史: 91 y/o man discharged from CM division on 1130912 due to 1. Community acquired pneumonia over RUL s/p Unasyn(113/9/8-9/12) 2. Hyponatremia, suspect poor intake or HF related s/p N.S. 500c.c. QD and OPD diuretics therapy 3. Hyperkalemia, r/o CKD or medication related s/p Kalmate use(9/8-9/9) 3. hypertension 4. CKD; no allergy hx; TOCC(-; AZ2-moderna2); He is brought to ED because of fever, dry cough, and dyspnea this morning. There is no chest pain, no headache, no neck pain, no focal weakness, no legs pain/edema, no cellulitis, no arthritis, no abdominal pain, no dysuria.					
個人病史: quit smoking					
過去病史: HTN (Hypertension)-(s/p PPM, 113/03/01 2D: +--+AR, +MR/TR, EF 65%);DM (Diabetes mellitus);CKD - Chronic Kidney Disease-(baseline Cr:1.5-1.6);COPD-(OPD: Seretide+Spiriva);					
物理檢查 DATE:1130924 TIME: 1149 TRAUMA.N					
生命徵象 1130924 BP:115/58mmHg,HR:83beats/min,RR:18times/min,BT:37.8°C					
GENERAL CONDITION Ill-looking appearance					
MENTALITY GCS:E4V5M6;					
Clear consciousness; intact orientation					
HEENT No pale conjunctiva, no icteric sclera; pupil size and LR: R/L: 3+/-3+/-					
NECK Supple neck; no palpable lymph node					
CHEST & LUNGS: Clear breath sound; no rales, no wheezing					
HEART Regular heart beat; no murmur					
ABDOMEN No operation scar, normoactive bowel sound, no tenderness, no Murphy's sign, no McBurney's point tenderness; no rebounding pain					
BACK & SPINE: No CV angle knocking pain					
EXTREMITIES No limitation of movement; no pitting edema					
EXOGENITALIA					
RECTODIGITAL EXAM					
NEUROLOGICAL EXAM					
CRANIAL N.: intact					
SENSORY MOTOR: Muscle power (limbs): RUL/LUL: 5/5; RLL/LLL: 5/5					
REFLEXES Babinski sign: R/L: plantarflexion/plantarflexion					
GAIT					
OTHERS 皆微弱或功能急遽下降，請以CFS 臨床衰弱量表評分(在 POMR 選項)，有利病人收治高齡病					
房					
負責主治醫師:62381王中邦 負責醫師:王中邦 助理醫師:62381王中邦					
編輯 列印 關閉					
已電子簽章(簽章序號0000000319122059)					



PrivNurse

Summary

Patients

Management

Logout

PrivNurse AI Assistant System

Medical Record No.

MR202508012

Patient Category

NHI GENERAL

Patient Info

David Chen(M) 72.00kg

Department

Orthopedics

Birthday

1952-03-15 (73y 5m 30d)

Admission Time

2025-05-17 (14:30:00)

Bed Number

C715-2

Status

HOSPITALIZED

Discharge Note

Consultation

Nursing Note (STT)

Chief Complaint

Post-operative wound infection with cellulitis and abscess formation following left lateral malleolar fracture

Lab Report

Wed, Aug 6, 2025

Troponin I

上午08:00

0.02 ng/mL

Normal: <0.04

NORM

BNP

上午08:00

185 pg/mL

Normal: <100

HIGH

Mon, Aug 4, 2025

Wound Culture - Prevotella timonensis

上午08:00

Light growth

Normal: No growth

HIGH

Wound Culture - Finegoldia magna

上午08:00

Moderate growth

Normal: No growth

HIGH

Wound Culture - Bacteroides thetaiotaomicron

上午08:00

Heavy growth

Normal: No growth

HIGH

Diagnoses (6)

PRIMARY

2025/8/4

Cellulitis of left lower limb with abscess

SECONDARY

2025/8/4

Infection and inflammatory reaction due to internal fixation device of left fibula

PRESENT

2025/5/17

Displaced fracture of lateral malleolus of left

Treatment Course

Generate AI Summary

Treatment course will be generated here, or you can enter manually...

Specialist Consultation

[2025-08-06T21:29:33]

病患因蜂窩性組織炎併發膿瘍形成，傷口培養顯示有 Bacteroides thetaiotaomicron、Finegoldia magna 和 Prevotella timonensis 等多種厭氧菌感染，目前使用 Curam 和 Metronidazole 治療，申請感染科會診以獲取進一步的感染診斷和治療建議。感染科 Dr. Zhang 回覆建議使用 Subacillin 1.5 gm iv Q8h 或 Curam (1 gm) 1# bid po + Metrozole (250) 1# qid po。

Nursing Note Summary

[2025-08-06T21:50:55]

VitalSign: Asian tolerated cardiac medication well. Blood pressure stable.

[2025-08-06T21:45:35]

Objective: Patient mentioned family history of heart disease.

[2025-08-06T08:00:00]

Evaluation: Morning assessment: Significant improvement noted. Wound drainage minimal, erythema reduced to 5×4cm.

[2025-08-06T07:00:00]



LLM於心臟超音波翻譯摘要

解決問題:英文心臟超音波報告量大且異質，跨科溝通與病人可讀性不足。

運用AI技術: 利用 Fine-Tuning 後的 LLaMA 翻譯和摘要心臟超音波報告 (雙模型)

單位合作: 台中榮民總醫院

- 訓練設備: Tesla T4 16GB
- 部屬設備: RTX 5070 Ti
- 基礎模型: LLaMA 3.2 3B Instruct



The screenshot displays the LLaMA web interface with the following settings:

- 設定 (Settings):**
 - 翻譯模型 (Translation Model):** llama-3.2-3b-instruct
 - 摘要模型 (Summarization Model):** llama-3.2-3b-instruct-summarizer-q5
 - 風格 (Style):** Clinical
 - 保留段落/清單結構 (Keep paragraphs/list structure):** ☒
 - 翻譯 (Translation):** max_tokens / temperature: 2048 / 0.12
 - 摘要 (Summarization):** max_tokens / temperature: 512 / 0.20
 - top_p (0-1):** 0.90
- 英文原文 (English Original):** 貼上固定格式的心臟英文資料 (Paste fixed format cardiac English data)
- 翻譯結果 (繁體中文) (Translation Result (Traditional Chinese))** and **摘要結果 (Summarization Result)** buttons.

LLM於心臟超音波翻譯摘要

- 研究目的 (核心目標) :

1. 開發**容器化臨床文字處理系統**，實現**雙階段翻譯與摘要流程**。
2. 運用**大型語言模型與高效微調技術 (LoRA / QLoRA)**，支援**本地端低資源部署**。
3. 建立**數值 / 單位一致性檢查機制**，降低病歷生成過程的偏差風險。
4. 提供模型與參數即時切換，便於**多模型比較與臨床 / 教學示範**。

- 研究重要性 (價值與貢獻) :

1. 操作彈性：支援**模型與參數即時切換**，便於比較與測試。
2. 數據真實：導入**數值 / 單位檢查**，避免臨床數據偏差。
3. 臨床價值：心臟超音波報告翻譯後仍艱澀難懂，提供**簡單易懂友善翻譯和摘要**。



系統登入

輸入密碼

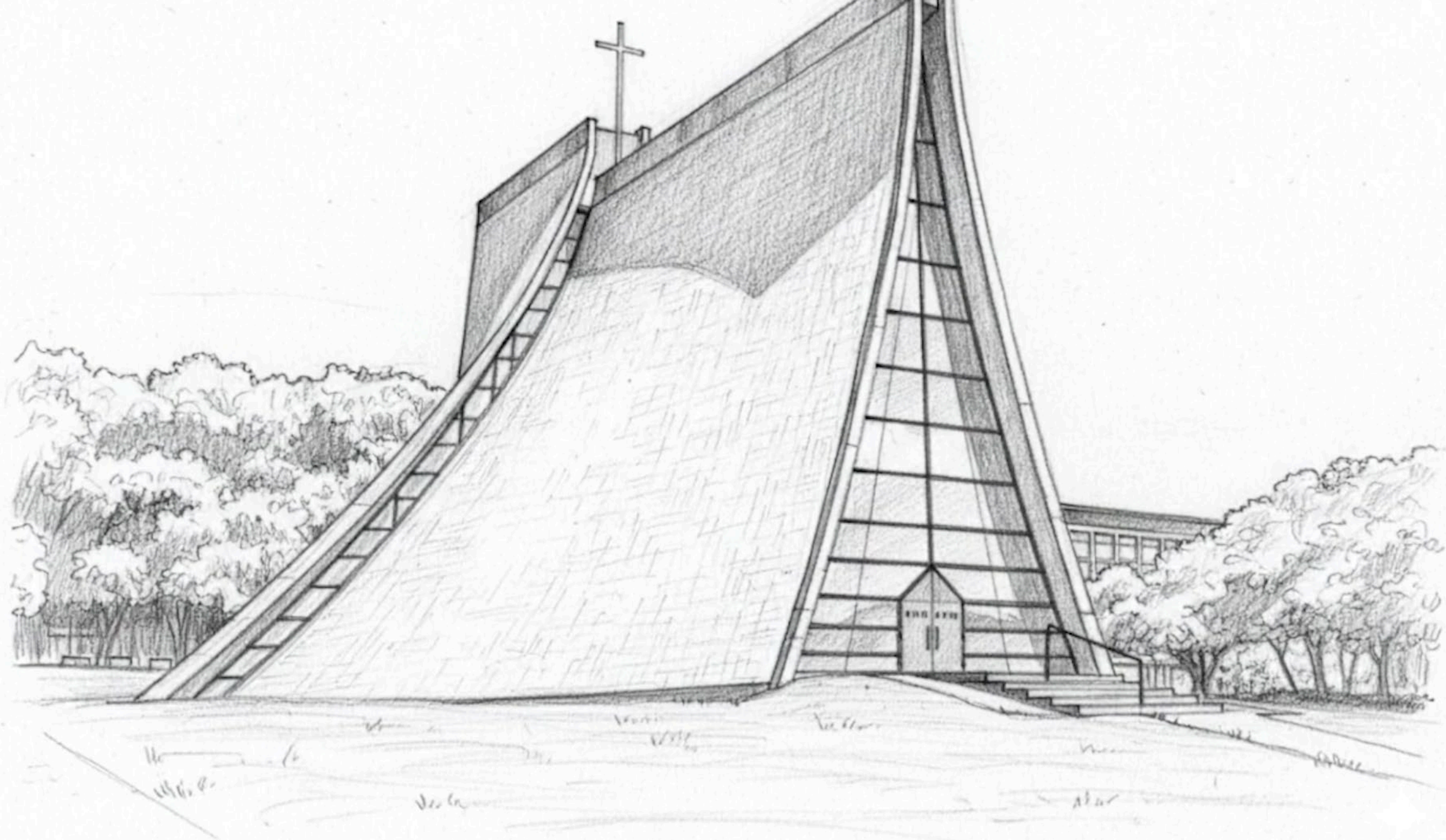
大家好


此為臨床診斷報告書摘要和翻譯文字系統的Demo



胸痛病患臉部表情辨識系統及其方法

系統展示操作





東方學府育英才
海闊天空志不改
慧詰創新展光彩
與時俱進拓新脈
生生不息如花開
成功卓越創未來

感謝聆聽

Thank You

Speaker contact information

