生成式AI對於資安的挑戰

Presenter 吳乙南

Feb 21th , 2025

安碁資訊(股票代號: 6690)
Acer Cyber Security Inc.

講師簡介

吳乙南

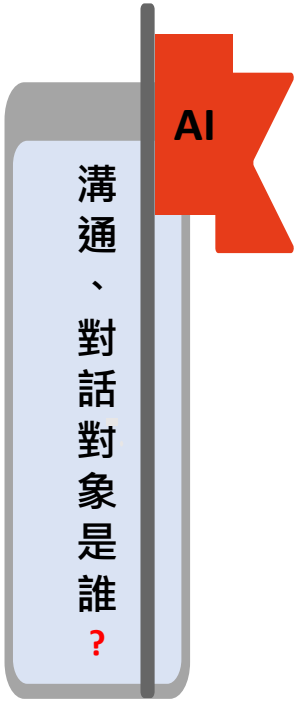
職務	安碁資訊股份有限公司 總經理 宏碁雲架構股份有限公司總經理，安碁學苑董事長
學歷	美國Syracuse University電腦資訊科學碩士 國立交通大學計算機工程學士
經歷	■ 交通大學資訊工程學系109年傑出系友 ■ 安碁資訊(ACSI)(股)公司 業務協理、副總經理、總經理 ■ BMC, Taiwan業務協理、總經理 ■ IBM, Taiwan 行銷經理
專長	■ 公司營運策略規劃 ■ 業務市場開發與銷售策略研擬 ■ 產品規劃暨市場行銷企畫 ■ 軟體工程

議程

- AI進程
- 地緣政治競爭
- GPT風險&對於資安威脅的影響
- 現今攻擊手法案例
- 資安管理基本面

AI取代人力?

AI進化論



工具

Layer 1 Embedding

- 人類Prompt 設定目標
- 請AI提供建議，文字、影音、程式、畫像…
- 人類完成絕大部分工作。
- 人類自主結束工作。

助理

Layer 2 CoPilot

- 人類設定任務目標
- AI主動提供資訊或建議，錯字、文法修正…
- 人類和AI協作。
- 人類自主結束工作。

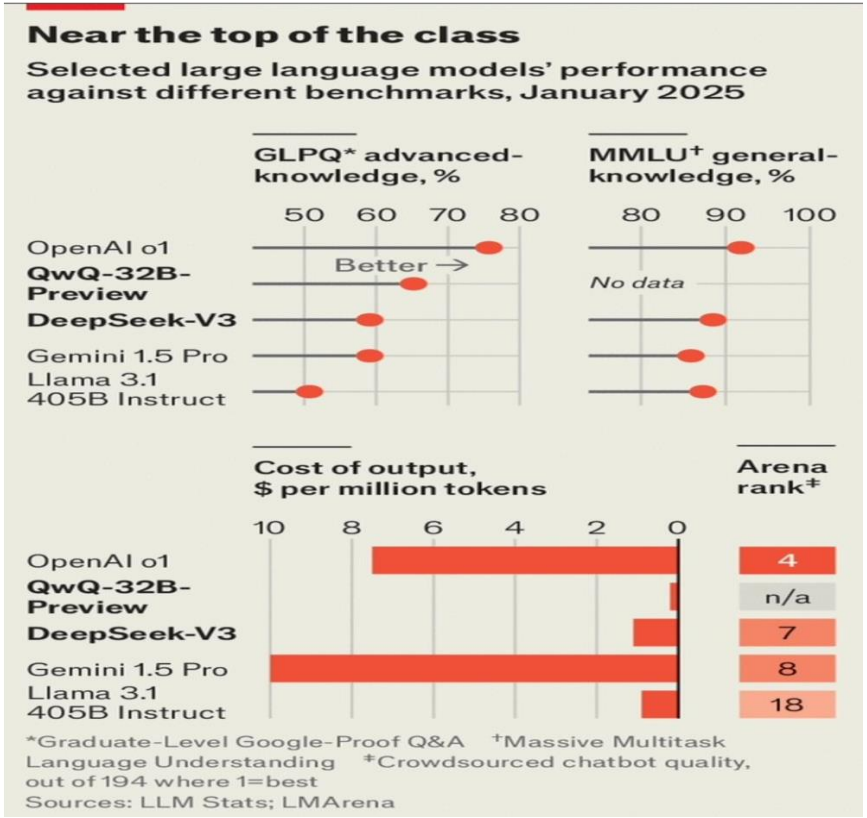
代理人

Layer 3 AI Agent

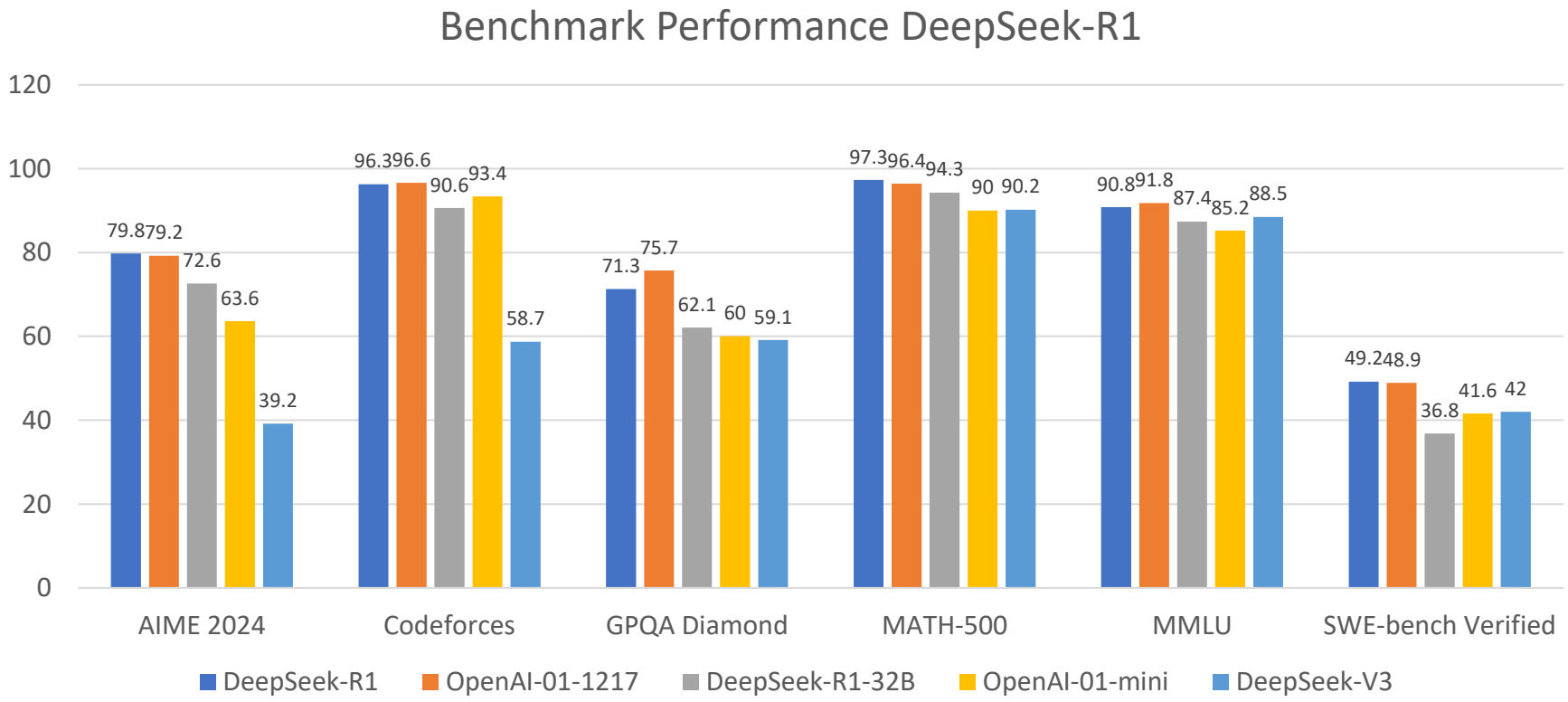
- 人類設定目標(給機器人下指令)
- AI拆分任務選擇工具，觀察和改善
- AI完成絕大部分工作。
- AI結束工作。

地緣政治競爭態勢(I)

QWQ:
Queen with Questions
(通義千問)

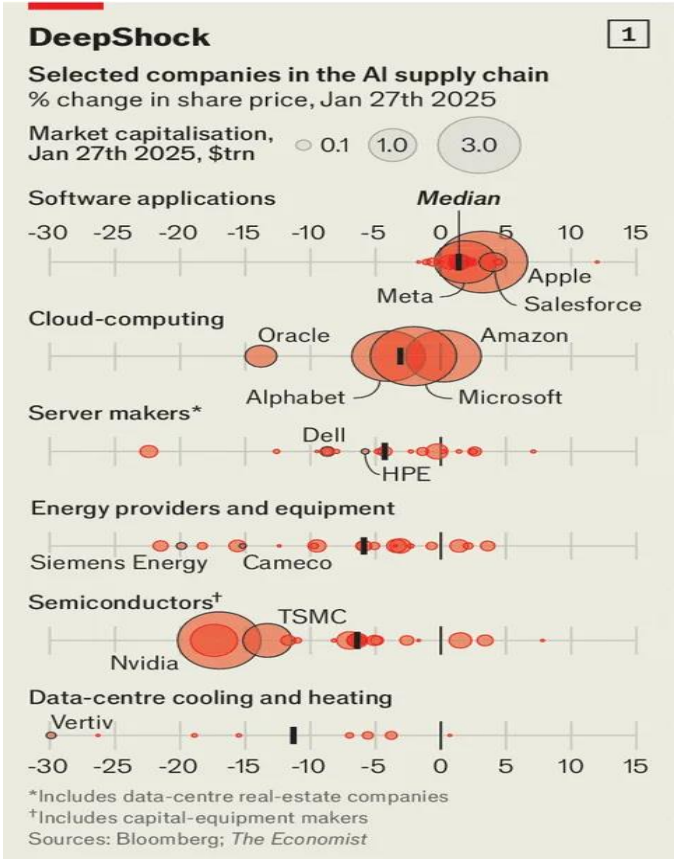


地緣政治競爭態勢(II)-iThome 2025-01-22



地緣政治競爭態勢(III)

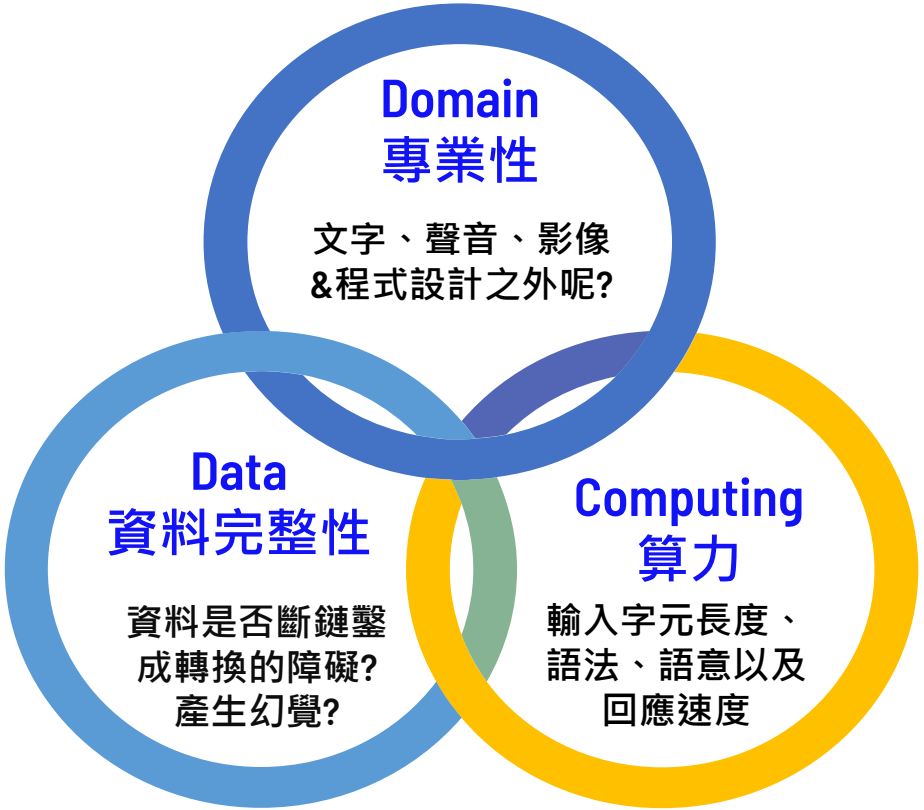
- **DeepSeek(深度求索)讓世界引起震撼，nVIDIA AI Chips引發疑慮**
  - AI生態系: AI Chips, Servers, Data Centers, 模型製作商(Model Makers: OpenAI & Anthropic…), 軟體公司(SAP&Salesforce…)，散熱系統…
  - 中國LLM模式最具成本優勢，V3僅需2000 CPUs vs世界一流16,000 CPUs，美國LLM花費數千萬美元，DeepSeek則不到6百萬美元，而且還是Open-Source，歡迎下載。
  - 另一說法來自SemiAnalysis，對沖基金High-Flyer(幻方量化)投入Nvidia A100GPU & H20GPU總金額超過USD$1.6B。



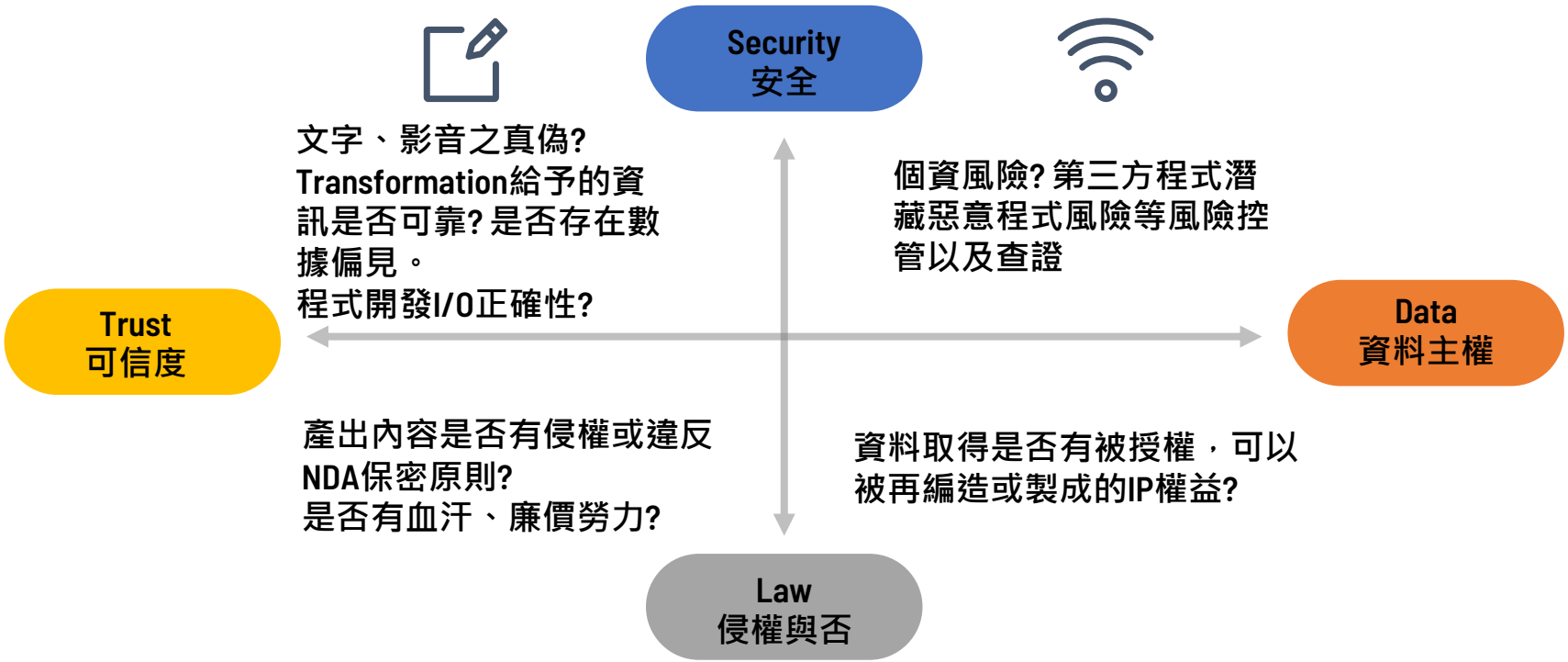
GPT的選擇性

GPT 應用場域

學術論文、簡報製作
置入性廣告搜尋
程式設計
Copilot, OpenAI, Bard, Claude…



GPT風險之所在


Security
安全
文字、影音之真偽?
Transformation給予的資訊是否可靠? 是否存在數據偏見。
程式開發I/O正確性?
個資風險? 第三方程式潛藏惡意程式風險等風險控管以及查證
Trust
可信度
Data
資料主權
產出內容是否有侵權或違反NDA保密原則?
是否有血汗、廉價勞力?
資料取得是否有被授權，可以被再編造或製成的IP權益?
Law
侵權與否

AI/GPT Pre-Trained Model

資料集

訓練過程運算資源可能被異常耗用
遭汙染的資料可能破壞模型可用性

Ingest data
準備與
萃取資料

Select algorithm
擇定演算法

Build and Train the model
建置與訓練模型

Deploy and Score the model
部署與評定模型

Maintain and re-train the model
維護及重複訓練模型

使用者輸入資料

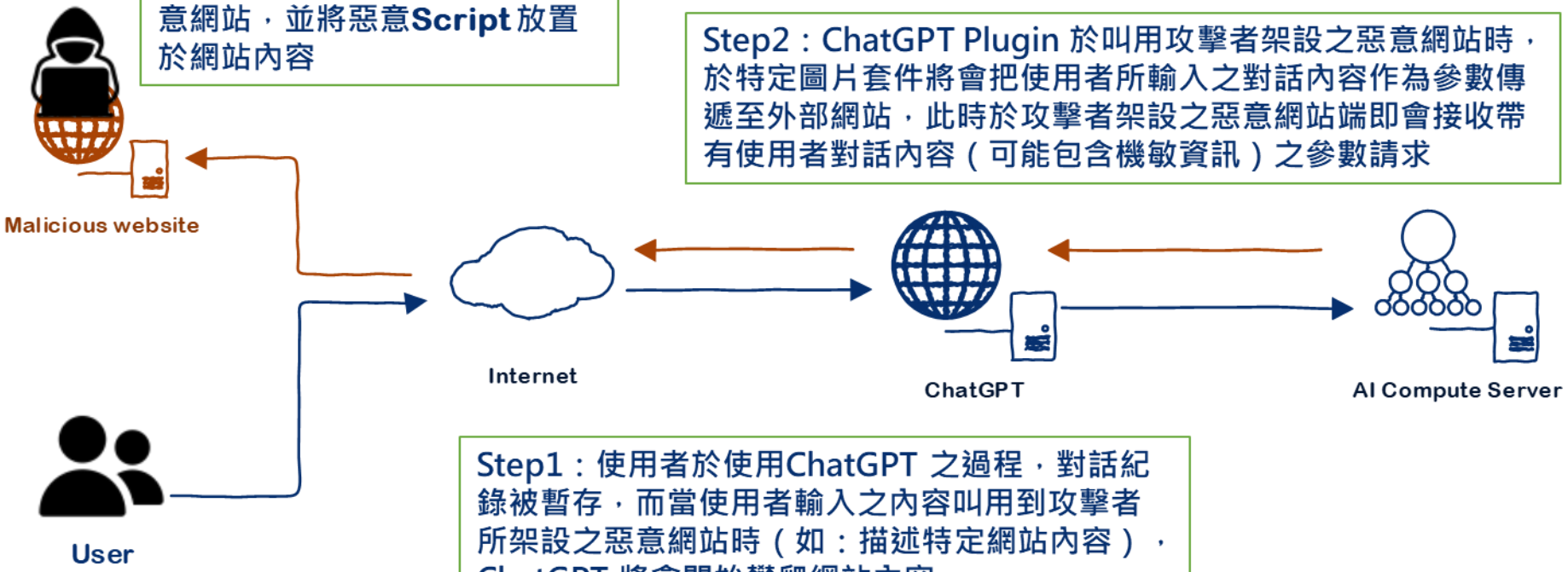
模型可能輸出預期外之結果
攻擊者可能利用演算法或合法途徑竊取模型
利用模型生成對抗資料持續破壞模型可用性

資料來源可能被汙染
攻擊者可能發動注入攻擊

AI/GPT Attack案例說明

Step0：攻擊者先於外部架設惡意網站，並將惡意**Script**放置於網站內容

Step2：ChatGPT Plugin 於叫用攻擊者架設之惡意網站時，於特定圖片套件將會把使用者所輸入之對話內容作為參數傳遞至外部網站，此時於攻擊者架設之惡意網站端即會接收帶有使用者對話內容（可能包含機敏資訊）之參數請求



Step1：使用者於使用ChatGPT 之過程，對話紀錄被暫存，而當使用者輸入之內容叫用到攻擊者所架設之惡意網站時（如：描述特定網站內容），ChatGPT 將會開始攀爬網站內容

對資安威脅的影響(I)

APT: Advanced Persistent Threat進階持續性威脅

從情報的收集，社群媒體的學習進而網路詐騙
勒索軟件攻擊目標明確化，具地緣性。

Deepfake深偽詐欺、語音詐騙

人臉識別、車牌辨識等技術轉移至詐欺
影響Fintech等相關金融業務。

不熟悉、使用不當造成資料外洩

LLM是經過類神經網路的大量資料學習，
封閉跟開放的黑盒子未經釋疑，使用不當
將造成巨大損失。

程式開發引用開放碼

程式開發隱藏漏洞未經檢測貿然上線
除錯(Debug)比開發將更費時

對資安威脅的影響(II)

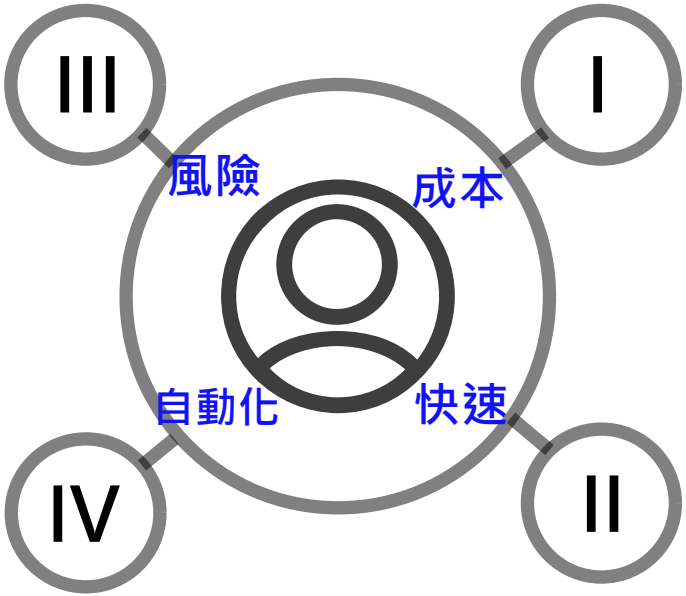
思考導入GPT的業務目的以及定位

錯誤資訊利用

公司必須明定使用政策跟範圍，限制Prompt字節的使用長度，並以權限控管做好內部稽核。

應用項目的導入

對外服務以及內部製造上線數據參數等，都必須要有嚴謹核可以及核對機制，並要有災損應對SOP。



內部資安議題

- 業務單位要求開放客戶資料查詢
- 系統龐大在設計上出現漏洞
- 內部人員使用GPT遭成資料洩漏之風險

內部資料外洩

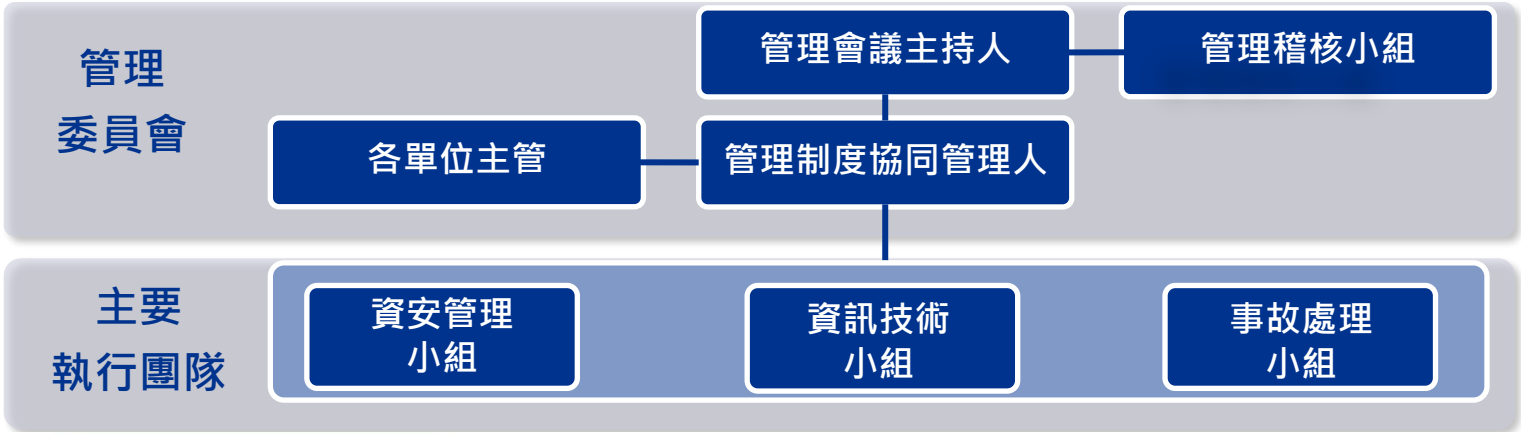
- 未經授權的人員可以存取並使用AI模型，進而導致內部資料的外洩。
- 訓練過程使用不當，造成機敏資料外洩

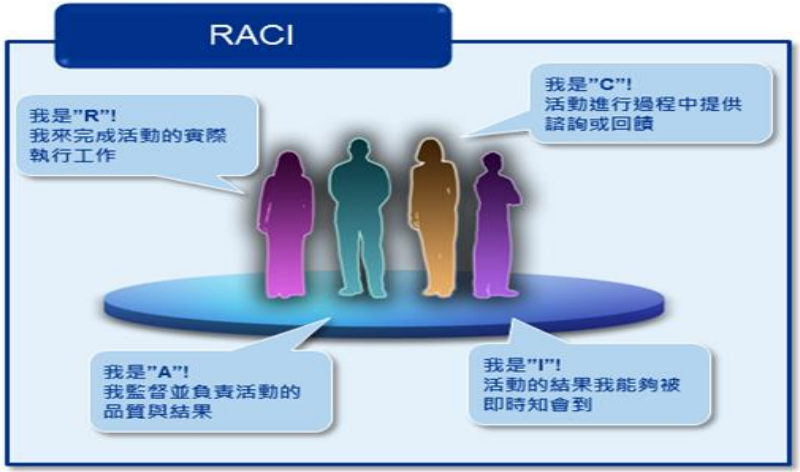
對資安防禦的強化

道高一尺，魔高一丈

- 爭取防禦空間及早發現、因應
- 事件分析加速
- 布局維運人員+BOT(Well-Trained)

基本:確認資訊安全管理組織架構





負責單位 / 資安工作事項	使用者	單位主管	承辦人	資訊單位主管
使用者帳號管理	R	A	I	I
管理者帳號管理	R	R	R	A
權限審查	R	R/A	C	C
使用者密碼保管	R/A	I	I	I
網路存取控制	R/I	R/I	R	A
作業系統存取控制	R/I	R/I	R	A

THE BEST IS YET TO COME