

出國報告(出國類別：其他)

出席第七屆  
首爾人工智慧高峰會國際研討會  
(AI Summit Seoul 2024)

出國人：

服務機關：國家科學及技術委員會

姓名職稱：周子元處長、陳李賜楠科長

派赴國家：首爾，韓國

出國期間：113年12月9日至113年12月12日

報告日期：114年1月7日

# 摘要

本次研討會為期 2 天，研討會內容除規劃介紹人工智慧（AI）的發展；講者分享進行 AI 投資與應用時，需考慮的重要要素與導入建置所屬 AI 模型時應該注意的事項，並提及可能會面臨的挑戰與機遇有哪些項目；為讓與會者更有所感，大會安排多位講者針對各 AI 應用領域分享該公司 AI 導入與應用實戰經驗、對未來發展趨勢進行分析，針對實際 AI 產品進行相關預錄影片播放介紹，強調以人為本的 AI 系統開發；部分講者建議 AI 產品發展策略應朝模化方向邁進、企業藉由 AI 轉型邁向智能企業、數位平台公司藉由 AI 代理進行轉型、多場會議內容更針對大/小語言模型(S/LLM)進行分析比較與介紹適合導入之場域，整體內容堪稱扎實。相關各場次研討會內容摘要，請詳見報告內容，僅提列幾項重點如下：

- (一) AI 初創公司的估值高漲，非完全基於產品實際價值，而是因基礎設施和訓練成本非常高，導致估值上升。AI 領域數據和計算資源需求極大，需大量的資金投入，如：數據獲取和授權須支付大量費用，曾經為免費使用的數據來源，未來可能須支付價金。這是身為 AI 投資者，其在預期回報上須予以考量的挑戰(或有大量投資流入，銷售額仍然低之現象)。
- (二) 企業應專注於分析現有解決方案，並確定哪些解決方案最適合自身的需求。如此可避免在自建模型上浪費過多資源。最實際的做法是設立標準，並通過結合公共經濟和研發來推動 AI 技術的發展。即應該在現有的技術基礎上進行合作，而非單打獨鬥。特別是在面對像 LLM 這樣的高成本技術時，依賴於已有的解決方案可能會是一個更具成本效益的選擇。
- (三) AI 的發展應該與人類的需求相對接，包括提供創造性表達的支持、增強社會互動、促進個人責任感等。作為設計師和開發者，目標應該是創建能夠支持自我提升的 AI 系統，此系統能夠幫助使用者增強自身能力，而不是取代他們。類似於相機、導航等工具，AI 應是用來增強創造力、促進社交，而非削弱人類的核心價值。
- (四) 隨著物聯網等新技術的普及，AI 的應用將不再局限於傳統領域，而是深入到更多的行業和業務流程中。即 AI 技術未來將改變各行各業的運作方式，尤其在金融和風險管理領域更甚，如何利用 AI 來提升運營效率、客戶體驗及解決技術挑戰，是成功者的關鍵因素。
- (五) 在企業內部，要觸發 AI 廣泛應用，員工需要具備一定的創新能力和靈活應變的能力(非僅是技術專家)。因此，AI 的成功應用不僅需要技術的支持，還需要企業文化的變革。隨著 AI 技術的發展，未來工作模式將發生根本變化。許多職位或可能消失，但同時新職位和角色會出現。這些新角色將要求員工具有更高的專業技能，且工作內容變得更加專業化和細化。
- (六) AI 系統在應用過程中出現錯誤或偏差時，誰應該負責？此不僅關於個人責任，還涉及數據創建者、學習者和使用者的責任區分。針對 AI 在成長過程中所面臨隱私、泡沫化、偏見和道德等問題的挑戰，似應預為規劃因應。

# 目錄

壹、目的 .....	1
貳、過程 .....	2
參、參訪 Purestorage 首爾辦公室 .....	2
肆、研討會摘要 .....	5
一、人工智慧 (AI) 的發展、投資及應用面臨的挑戰與機遇 .....	5
二、以人為本的 AI 系統開發 .....	7
三、AI 產品的規模化(AI 產品發展策略).....	8
四、AI 轉型邁向智能企業.....	13
五、數位平台公司轉型(AI 代理) .....	16
六、AI 應用領域.....	16
1、生成式 AI 與搜尋介面的評估—以 Naverv 系統為例.....	16
2、視覺創作.....	17
3、人形機器人.....	18
4、LG AI 代理系統設計與發展.....	19
七、硬體設備(半導體)需求.....	20
八、語言模型(LLM).....	22
1、LLM 模型的測試與評量.....	22
2、小型語言模型與神經網絡.....	23
3、大型語言模型 (LLMs) 微調 .....	24
4、優化封閉網路 LLM .....	26
九、AI 代理來協助 AI 產品開發.....	28
伍、心得及建議 .....	29
一、心得.....	29
二、建議.....	30
陸、附錄：研討會議程及資料	

## 壹、目的

人工智慧(AI)近年如火如荼幾乎是以火箭上升的速度蓬勃發展，導致各企業/機關/組織沒有應用或導入相關 AI 模型或技術，就是落伍者的代表。因此，得知「麻省理工科技評論」(DMK Global) 規劃於韓國首爾舉辦 AI summit，期望藉由會議議程內容的安排，能幫助參與者更好地掌握整個價值鏈的變化，除提供全局也提供細節的全面視角，讓與會者能在最短時間內對 AI 的應用與發展更了解。因此，想藉由參與該次研討會來增長相關功力，作為未來國家科學及技術委員會(以下簡稱本會)導入建置相關 AI 應用或模型之參考。

本次研討會議程內容包括：探索最新的人工智慧技術並了解它們的應用，學習人工智慧、數據專家和研究人員的經驗(講者分享寶貴見解)。應用領域涵蓋金融、製藥、商業、製造和零售等行業，應用案例整合深度學習、機器學習、認知運算等技術。

## 貳、過程

本次參訪及研討會自 2024 年 12 月 9 日至 5 月 11 日止，共計三天行程如下：

日期	上午	下午
12/7(星期六)		桃園機場搭乘中華航空前往仁川機場
12/9(星期一)	參訪 Pure storage Seoul office	
12/10(星期二)	研討會開始	
12/11(星期三)	研討會開始	
12/12(星期四)	仁川機場搭乘中華航空前往桃園機場	

## 參、參訪 purestorage 首爾辦公室



圖一、贈送接待人員禮品(台灣的高山茶)

12月9日前往亞歐會議大樓 (ASEM Tower)，參訪 purestorage 辦公室，PureStorage 技術定位為全快閃儲存供應商。但從 112 年起產品組合更加多樣化，不再僅限於儲存產品，期望轉型成為平台公司，而不再是單純的儲存供應商。基於此，該公司產品組

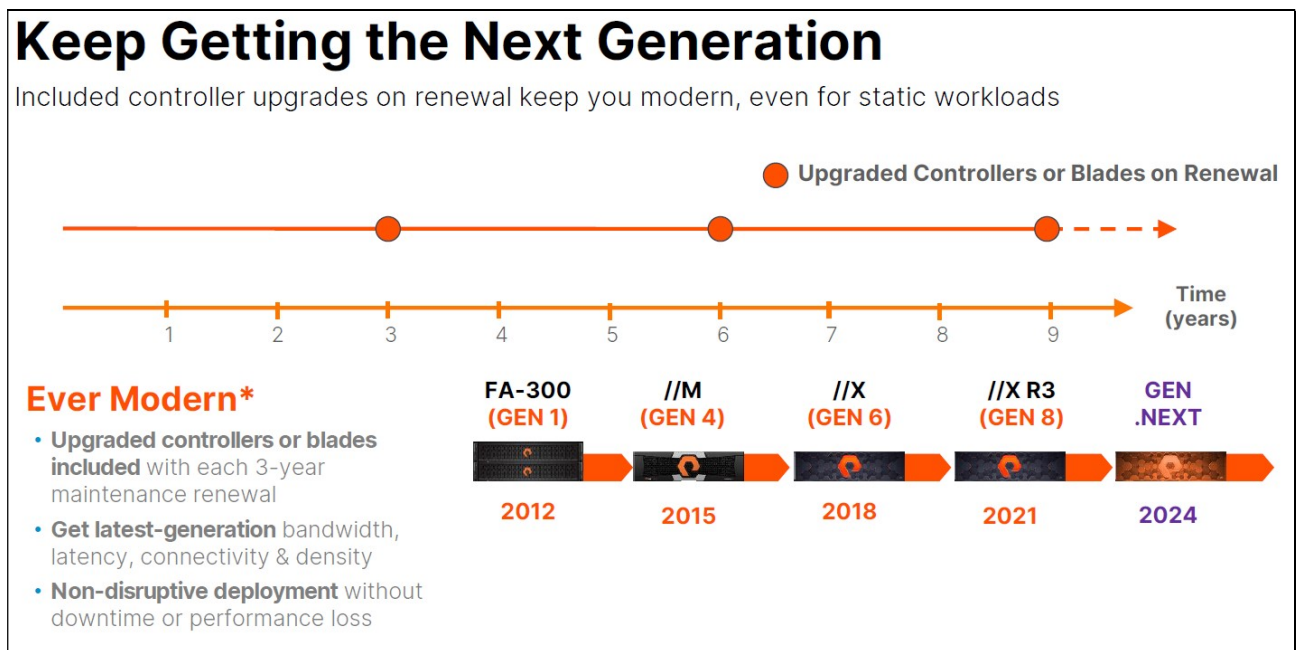
合涵蓋廣泛，不僅限於檔案和區塊儲存，還能支援容器化應用程式。此外，同時提供多種軟體和硬體解決方案。基於產品組合我們正在推動各種基礎架構，不僅限於在地端（on-premise），也能針對公有雲或其他服務，向客戶提供我們的產品組合。

Purestorage 指派 YS Kim 經理接待，面對面互動的討論經費、維護授權(MA)、設備汰換等問題交換意見，並簡報產品平台特色。

### 一、Evergreen 架構。

Evergreen 架構有兩個核心特色：1.無破壞升級，客戶在升級至下一代型號或版本時，不需要停機或進行資料遷移。2.持續升級，如果客戶使用我們的 Evergreen 訂閱，它是一種維護續約模式。在這種模式下，客戶每三年都可以升級到最新的型號，且不需要進行資料遷移或服務停機。

基於 CPU 架構的發展，通常每兩年就會推出新一代的處理器。在這種情況下，我們需要採用新的 CPU 或記憶體技術並進行開發與測試，然後在三年後正式推出新的型號。PURESTORAGE 將硬體與軟體進行了分離。在這個架構中，硬體只負責資料平面，而軟體負責控制平面。所以當客戶想要升級到更高階或下一代的控制器時，只需在線上進行替換，無需停機或遷移資料。硬體被替換後，系統立即識別並完成升級。



圖二、硬體設備續約模式

所有的產品都是基於模組化系統而設計的，因此，無論是控制器模組、磁碟模組，或硬體模組，客戶都可以在不中斷服務的情況下隨時進行替換或升級。這就是第一個架構特色。第二個特色是我們提供長壽命的機箱，機箱是一種基礎硬體。如果機箱故障，客戶通常會面臨服務停機的需求，因為它是單一硬體輸入電源的核心元件，且因設計長壽命機箱，因此在升級至下一代產品時，機箱無需更

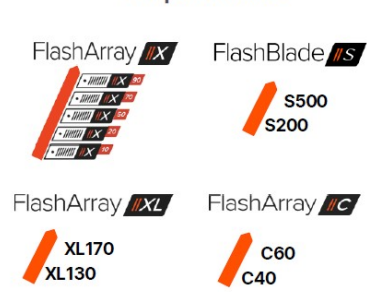


換。我們規劃的機箱壽命高達 10 年，只需更換控制器或磁碟模組即可升級至下一代型號。這就是能夠在不中斷服務的情況下進行一切更換和升級的原因。此外，所有模組都可以進行升級與替換，這完全符合 Evergreen 架構的理念。換句話說，客戶只需購買一套陣列，就可以使用 10 年而不會遇到服務終止的問題。每當推出新型號時，只需按三年一度的訂閱模式付費，即可將所有模組升級至最新版本。但是有一個條件，那就是相同的合約期限。例如，購買我們的儲存設備，並選擇一年的保固，我會接受，然後明年如果你想繼續保一年，我們不會增加 MA 費用，然後如果第二年我們想再延續一年，我們也不會增加，因為條件不變，只要你保持相同的條件，我們從不會對客戶增加維護費用。

## Evergreen: It All Starts With Architecture

The modular, software-defined, renewable architecture of FlashArray & FlashBlade//S


**Upgradable  
Controllers & Blades**  
For performance



FlashArray **XN**    FlashBlade **//S**  
S500  
S200


FlashArray **//XL**    FlashArray **//C**  
XL170  
XL130    C60  
C40

**Long-life  
Chassis**  
10+ year lifespan



FlashArray  
FlashBlade **//S**

**Upgradable, Expandable  
Flash**  
For capacity & expansion



3D TLC, QLC, NVME  
Flash Modules

**Upgradable Software**  
Purity **FA**    Purity **FB**

**Upgrade Everything While Online**  
With no downtime, performance loss, or data migration

圖三、purestorage 系列產品

Purestorage 主要產品，FlashArray 是用於統一的區塊儲存和檔案儲存。有四個 FlashArray 的型號，分別是 XN、XL、C 和 E。XN 和 XL 是低延遲型號。如果你運行的是一些業務關鍵型數據庫，或是需要低延遲的應用程式，那麼在這種情況下，XN 或 XL 型號可選擇。所以如果有多個 XN 型號，你可以將它們整合到一個單一的 XL 平台上。從延遲角度來看，這些型號是相同的，當然。而 C 型號則是高密度和高效能的。這是我們產品中性能最好的型號，也是第二級產品的選擇。所以 FlashArray C 型號主要用於混合配置。(如與 IBM Storage 作為混合配置)，那麼在這種情況下，你可以選擇 C 型或 E 型。那 E 型是最高效能的。所以 75TB 的 FDM (Flash Drive Module) 主要針對更高容量，並且是針對硬碟陣列的需求。這些就是我們 FlashArray 的產品組合。而 FlashArray 是一個擴展式平台。所以如果

你運行的是任何 AI 工作負載或數據分析，在這種情況下，你可以使用 FlashArray。S 型則是超高效能或超高性能的產品，適用於數據分析。所以如果是 AI 或數據分析的工作負載，在這種情況下，可以使用 FlashBlade S 型。E 型則是更加高效的型號。適用於大容量、備份或針對硬碟驅動器的需求，這種情況下，可以選擇使用 FlashBlade E 型。

在設備上提供 Pure One 及 Pure Fusion 2 種軟體，Pure One 則是總體監控軟體。如果開通防火牆，並允許我們存取陣列的診斷日誌。只要你開放端口(PORT)，就會收集所有的陣列負載資訊，並將其傳送到公共雲端。接著，我們會監控並分析這些負載，甚至模擬你的陣列，並加入更多的程式碼來了解可能發生的情況。是所有基於 AI 的單一監控解決方案，如果採購任何一款產品或其他產品，你都可以使用 Pure One 平台來監控或模擬你的環境。這就像是 SOC（安全運營中心）一樣。我們可以連接這些資訊，或將其包裝成我們的單一監控儀表板。Pure Fusion 就是我們所有產品的統一控制平面。通過這個 Pure Fusion，並且將工作負載從一個陣列移動到另一個陣列，這一切都可以透明地進行，且不會有任何服務中斷的問題。如果您的工作負載有一些戰略性位置，那麼您可以在各個陣列之間平衡所有的工作負載。因此，Pure One 和 Pure Fusion 是單一的 control 平面平台，軟體、API 等工具。

PureStorage 的產品和解決方案具備高性能、可擴展性和全面的監控支持，能夠滿足現代企業在 AI、數據分析和混合雲環境中的多樣需求，並以高客戶滿意度和市場領導地位為基礎，提供卓越的價值和支持。

## 肆、研討會摘要

### 一、人工智慧（AI）的發展、投資及應用面臨的挑戰與機遇。

#### 1. AI 的基礎設施與巨額成本

AI 基礎設施的巨額成本，如 OpenAI 預期的 70 億美元的成本，且根據統計，三分之一的投資，或者說每三家公司中就有一家，是 AI 初創公司。此比例遠高於預期，顯示 AI 投資的熱潮和潛力。AI 的增長速度和投資回報仍然值得關注，AI 領域的快速發展和投資機會，並為與會者提供了投資 AI 時需要考慮的要素。許多 AI 初創公司的估值高漲，並非完全基於產品的實際價值，而是由於基礎設施和訓練成本非常高。這導致估值的上升。AI 領域，數據和計算資源的需求極大，這需要大量的資金投入。例如，數據的獲取和授權現在需要支付大量費用，而曾經免費使用的數據來源變得更加昂貴。

儘管有大量投資流入，銷售額仍然低，尤其是非開放平台（如 LLM）公司的銷售額相對較低。這使得投資者在預期回報上面臨挑戰。投資者面臨選擇是投資於 AI 的底層基礎設施（如芯片、伺服器、數據標註）還是應用層。演講者選擇了應用層，因為許多像 Salesforce 和 Zoom 這樣的公司，雖然是基於基礎設施，但其成功來自於面向具體市場的應用。

全球的企業對 AI 投資存在顯著擔憂，特別是 AI 技術的開發和運營成本非常高。儘管 AI 具有巨大的潛力，但如果沒有足夠的知識和技術背景來有效運營，企業



可能會面臨資金上的巨大負擔。這讓企業在 AI 技術的投入上變得更加謹慎，AI 投資是否能夠帶來回報仍然是企業面臨的一個問題。對於大多數公司來說，建設像 LLM（大型語言模型）這樣的模型並不現實。相反，企業應該專注於分析現有的解決方案，並確定哪些解決方案最適合他們的需求。這樣可以避免在自建模型上浪費過多資源。最現實的做法是設立標準，並通過結合公共經濟和研發來推動 AI 技術的發展。這意味著，企業應該在現有的技術基礎上進行合作，而不是單打獨鬥。特別是在面對像 LLM 這樣的高成本技術時，依賴於已有的解決方案可能會是一個更具成本效益的選擇。

## 2. 語音助手與大型語言模型（LLM）

語音助手（如蘋果的 Siri、Google Assistant、亞馬遜 Alexa 等）自 2010 年起逐漸流行。這些助手主要依賴自然語言理解（NLU）來解析用戶語音並提供回應。蘋果 Siri 和 Google Assistant 在全球擁有超過 5 億用戶，亞馬遜 Alexa 也有相似的影響力。三星的 Bixby 等語音助手擁有數千萬的用戶，隨著大公司的投入，語音助手成為熱潮。語音助手的開發逐漸吸引了大量工程師，三星和其他大公司組建了規模龐大的語音助手開發團隊。語音助手的運營和增長與月活躍用戶（MAU）密切相關。儘管這些助手的投入金額巨大，並一直存在回報的疑問。

過去，語音助手主要用於簡單的任務，如設置鬧鐘和查詢天氣等，因此並未帶來顯著的收入。但隨著大規模語言模型（LLM）的出現，語音助手的能力得到了提升，開始能夠處理更複雜的任務，並引入付費服務。自 2022 年以來，LLM（如 GPT-4、GPT-5）引領了語音助手領域的革命。它們不僅能處理簡單任務，還能應對複雜的專業問題，如法律諮詢、費用計算等。LLM 的出現使語音助手能夠理解更加複雜的上下文，並能進行更高效的語音識別和語音合成，大大降低了延遲。未來，語音助手技術將進一步應用於亞洲市場以及機器人領域。語音助手將在自駕車、AI 教學助手等領域發揮更大作用。像 Aflo 這樣的多模態商業模式將成為未來的趨勢，語音助手將不再僅限於語音互動，還將結合其他技術（如設備端技術、機器人等）進行整合。。

## 3. AI 在製造業中的應用與挑戰

製造業中，AI 並非唯一的生產力工具，許多生產過程依賴於計算機視覺自動測試。AI 在製造領域的應用受到基礎設施（如 AI 基礎設施）和連接問題的影響而且 AI 在該領域的生產力提升難以立即證明。這些問題對 AI 在生產中的高效應用提出了挑戰。未來三年將在製造業中發生顯著變化，尤其是 AI 基礎設施的發展將大大推動該領域的進步。

各國將製造業視為戰略技術領域，並正加大對製造業的推動。AI 在製造業的應用將成為未來幾年技術發展的重點，這反映出 AI 對提升製造業生產力的潛力。AI 在提升生產力方面仍面臨難以量化的挑戰，基礎設施及連接問題是主要瓶頸。預計未來三年，AI 基礎設施的發展將顯著推動製造業的應用進步。

## 4. AI 的倫理與責任問題

AI 系統在應用過程中，當出現錯誤或偏差時，誰應該負責。這不僅僅是關於個人責任，還涉及數據創建者、學習者和使用者的責任區分。這提醒我們，AI 的責任問題比過去的技术更加複雜，需要從更理性和系統的角度來考慮。

當 AI 系統出錯時，如何準確區分各方的責任是至關重要的，並且這種區分對未來的發展將變得更加重要，尤其是在成本效益和可控性方面。在使用 AI（如大型語言模型 LLM）時，可能會遇到一些問題，這些問題是否能通過改變數據來修正仍然不確定，這反映出 AI 技術的可控性和修正能力仍然是未來發展中的挑戰。

會面臨像互聯網泡沫一樣的情況，AI 的基礎技術和商業模式的發展與互聯網類似，但 AI 的增長速度遠超過了互聯網。互聯網時代，從撥號到 DSL，再到超高速網絡，發展過程緩慢。而 AI 的崛起速度非常快，並且 AI 更集中在 B2B 和 B2C 的商業模式上，與互聯網初期的諸多問題類似，AI 在成長過程中也面臨隱私、偏見和道德等問題的挑戰。

現在從零開始建設 AI 解決方案在市場中仍然是可行的。他認為，企業可以在市場代理機構中選擇性地使用 AI 解決方案，例如像 Composites Group 或 WB 這樣的機構，這樣可以更靈活地運用現有技術。AI 技術的普及和使用可能需要 2-3 年的時間才能讓更多人習慣並有效地使用這些技術。具俊同意這一點，並認為，隨著時間的推移，人們會在心理上和技術上更加能夠使用 AI 技術。

## 二、以人為本的 AI 系統開發

「以人為本的 AI」的核心理念，不在於技術本身而是如何設計出能夠幫助我們達成目標、提升技能和能力的技術。以增強、賦能並提升人類的各種能力。人類擁有非凡的能力，而這些 AI 技術應該在各個方面提升我們的能力，而非取代我們的角色。

首先需遵循一套明確的過程和原則，在設計界面時，需要開發並進行多輪測試，從少數用戶群體開始，逐漸擴大到更多用戶群體進行驗證。設計以人為本的 AI 應該服務於全球需求，並且與聯合國可持續發展目標相對應。強調了如何利用技術服務人類需求，並幫助改善人類福祉。而成功的 AI 設計需要遵循一個經過驗證的過程，這樣能夠避免失敗的風險。例如韓國的產業在電子產品、汽車和其他商業產品中，已經在某種程度上遵循了這些設計原則，這些原則有助於產品的成功。生成式 AI 技術在過去兩年中迅速發展，並帶來了巨大的期望。然而，這些技術也引發了很多關於其潛在風險的擔憂，儘管如此，只要採取正確的技術與設計方法，這些風險是可以管理的，並且技術可以成功地服務於人類。

HCAI（Human-Centered AI）原則的核心在於設計支持人類價值的 AI 系統。透明度與控制是 HCAI 的關鍵概念。這意味著在提升自動化的同時，必須保持人類對系統的控制。這樣的設計確保了技術不會取代人類，而是增強、賦能使用者。以手機相機作為例子。現代手機相機擁有多種 AI 功能（如自動設置顏色平衡、焦距調整、降噪等），但是最終用戶仍然掌握控制權。用戶可以根據自己的需求進行調整、拍攝、編輯，甚至刪除照片。這種設計使得更多的人能夠拍出高質量的照片，並且他們感到自己擁有了更多的控制權和能力。

AI 設計必須始終以人類需求、安全和福祉為首要考量。AI 系統不僅應該服務於用戶的具體需求，還應該保護他們的隱私和安全，並尊重人類的價值觀。隨著技術的進步，未來的 AI 應該更加關注人類的需求與能力提升，並將賦能作為核心目標。設計出能夠提供高自動化的同時仍能讓用戶保持控制的技術，將會是未來 AI 成功的關鍵。AI 應該被視為一個超級工具，是一個強大的、增強用戶能力的工具，能夠幫助用戶做出更好的決策和提高效率，並保持用戶的主體性和控制權。AI 在決策過程中應該賦能使用者，讓

使用者能夠主動參與並控制整個過程，而不是讓 AI 代為決策。這樣的設計可以避免過度依賴技術，並確保技術成為提升人類能力的工具，而非取代人類的角色。

人類監督在 AI 系統中的重要性，尤其是在初期使用階段。新技術的引入需要高層次的人類監督，這是保證系統可靠、安全並值得信賴的關鍵。隨著技術的成熟，可能會逐步實現更高的自動化，但初期的人工監督 仍然是不可或缺的。根據時間框架的不同，人類監督 的形式會有所不同。短期內，監督可能需要以分鐘甚至幾小時為單位進行，而隨著時間的推移，當技術變得更成熟，監督的需求可能會減少。這種分階段的監督有助於平衡 AI 的自動化與人類的控制權。

設計以人為本的 AI 系統不僅是為了提高效率，更是為了確保系統不會被濫用，並且能夠增強用戶的 信任。透明度、可解釋性以及強有力的人類監督機制能夠確保技術的安全性與可靠性，並促進用戶對技術的信任。

AI 工具的驚人之處在於它們可以快速創建圖像、聲音等內容，但必須謹慎使用，特別是在處理複雜需求或高精度要求的情境中。能夠讓使用者在創作過程中有更多選擇和調整空間，從而避免 AI 系統生成過於偏離預期的結果。增強用戶控制權是推動 AI 系統向更人性化、以用戶為中心的方向發展的關鍵。生成假圖片、聲音和視頻，對民主和選舉造成威脅。生成型 AI 被用來進行大規模監控或隱私侵犯。未來可能被用於開發致命的自動化武器。取代工作仍然是一個被關注的議題，尤其是隨著某些行業的技術轉型，可能會出現一定程度的職位變動。生成型 AI 中的偏見問題和不公平性，以及由於不透明 和可解釋性不足造成的風險，依然是設計者必須關注的領域。歐盟的 AI 法案和美國的 AI 權利法案，指出政策正逐步推動更嚴格的監管框架來應對 AI 帶來的風險。應監管風險等級以確保高風險應用不會被濫用。儘管 AI 具有強大的能力，但人類對 AI 系統的責任是無可推卸的，AI 不能被賦予道德或法律責任。人類監督在 AI 系統中的作用不可忽視。特別是在關鍵性和高風險的領域，應該保持 強有力的人類監督，以避免系統失控或濫用

AI 的發展應該與人類的需求相對接，包括提供 創造性表達的支持、增強 社會互動、促進 個人責任感 等。作為設計師和開發者，我們的目標應該是創建能夠 支持自我提升 的 AI 系統，這樣的系統能夠幫助用戶增強自身能力，而不是取代他們。類似於相機、導航等工具，AI 應該是用來 增強創造力、促進社交，而非削弱人類的核心價值。

### 三、AI 產品的規模化(AI 產品發展策略)

AI 並不僅僅是技術的引入，而是涉及組織文化、業務流程、以及整體戰略的深層改變。大數據和 IoT 的目標往往是監控和收集數據。而 AI 的核心則是智能化決策，這意味著 AI 不僅是分析數據，它還能基於分析數據做出預測，或自動調整操作流程。因此，AI 的運用與傳統 IT 的區別，AI 不是簡單地提供數據支持或分析結果，它的目的是能夠基於數據進行深度學習，實現自動化和無人化的工作模式。

針對製造業的問題 AI 可運用於 AI 可以用來在原材料預測和檢查原材料的質量，提前剔除不合格的材料，從而避免質量問題蔓延到後續生產。這樣的預測性質量控制能顯著降低人工檢測的需求，並且提高生產效率。使用 AI 建立預測性維修模型，可以提前發現設備的潛在故障，從而避免生產中斷。這不僅減少了停機時間，也降低了維修成本。AI 技術能夠幫助企業逐步實現生產過程中的自動化，甚至無人化，這對於提升生產效率、降低人力成本具有重要意義。

硬體支持(如 GPU)在 AI 項目中的角色日益重要，因為 AI 模型需要海量的計算資源來處理數據，進行訓練和推理。AI 的引入不僅是一個技術問題，更是一個戰略問題。AI 能夠重塑企業的核心競爭力，並且決定未來企業的發展方向，企業的 AI 戰略應該包括 1. AI 的應用需要 IT 部門、業務部門和高層領導的密切合作，確保 AI 技術與企業實際需求的緊密對接。2. AI 的導入需要長期規劃，並且需要不斷的迭代與優化。企業不應僅僅關注短期的技術突破，還應該關注如何在未來的幾年中逐步拓展 AI 的應用範圍。3. AI 不僅僅是技術的導入，還涉及企業文化的轉變。4. 企業需要建立一種支持創新和持續學習的文化，這樣才能夠在 AI 時代中蓬勃發展。

1. AI 在顧客體驗創新中的作用，特別是對金融行業的影響，通常體現在以下幾個方面：

- (1) 即時數據處理和個性化體驗：這個平台將即時市場信息提供給 AI 代理，並能夠即時生成公司活動報告、測試結果以及市場信號的監控。這樣的即時反應對顧客來說，能夠提供高度個人化的服務。於證券交易中，根據即時市場變化提供的決策支持，可以使客戶即刻調整自己的投資策略，達到更精準的風險管理和收益最大化。

AI 在顧客體驗的創新中，能夠幫助企業通過預測模型來為客戶提供更個性化的建議。保險公司和銀行已經開始利用 AI 來生成動態的風險評估和保險建議，而在證券業，AI 可以用來分析市場走勢並即時給出投資建議，提升投資者的決策效率。

- (2) 增強的交互性和自動化服務：

AI 還能通過自動化客服和智能助手等技術提高顧客的互動體驗。通過語音助手或智能聊天機器人，顧客可以隨時隨地獲得信息和服務，這不僅提升了顧客滿意度，也減少了對人工客服的依賴。

2. AI 在提升內部員工生產力方面的作用：

- (1) AI 平台工具的創建：

Nielsen Zephyr 創建的“助手平台”通過這個平台，員工能夠將自己的工作需求與 AI 模型結合，從而自動化許多繁瑣的工作任務，這樣不僅能提升員工的工作效率，還能釋放出更多的創造力。這些工具的使用量遠超過預期，大約有 3200 名員工，其中一半以上的人都在使用這些工具。這證明了即便是非技術背景的員工，也能夠快速上手並利用 AI 提升自己的工作效率。這也讓員工從手動數據處理和反覆勞動性工作中解放出來，轉而專注於更高層次的決策和創新。

- (2) AI 在創建具體使用案例中的作用：

這些工具的成功關鍵之一，是其能夠促進員工創建和分享具體的使用案例。一些擁有多年開發經驗的員工，能夠為團隊成員創建 AI 代理，幫助他們完成具體工作任務。這樣不僅提高了整體團隊的工作效率，也促進了知識和經驗的共享，讓員工之間的協作更為高效。

此外，這種 AI 平台還能夠幫助公司發現隱藏的工作流程問題，並通過 AI 自動化進行優化。比如，在團隊中，經常會有人提出“為什麼要這麼做”的問題，通過 AI，這些疑問可以快速得到解答，並提出更加優化的工作流程。

- (3) AI 助力員工決策與工作流程優化：

AI 不僅能幫助員工完成特定的任務，還能協助他們在決策過程中提供支持。比如在金融市場中，AI 能夠通過算法預測市場走勢，並為員工提供投資策略建議。員工在面對複雜數據時不再是單純依賴直覺，而是基於數據支持做出更明智的決策。結合

AI 的工作流程優化也有助於提升生產力，尤其是在處理繁瑣的日常工作时，AI 能夠自動化許多手動流程，減少了員工的工作壓力並提高了效率。

3. 許多關於 AI 應用與策略的討論都觸及了當前和未來企業如何應對 AI 轉型的挑戰。

#### (1) AI 引入與應用的挑戰

數據與安全性問題是許多企業在實施 AI 時首先面臨的問題。特別是在金融和能源等敏感領域，數據的保護和隱私問題是首要關注的事項。許多公司選擇外部雲解決方案（如 AWS、Azure）來解決這些問題，並進行適當的加密和保護措施。

內部解決方案和外部解決方案（如 OpenAI）的選擇各有利弊。內部開發雖然能夠保持更高的控制權和安全性，但開發和維護的成本會相對較高，且可能無法跟上快速發展的技術變革。因此，許多企業選擇與外部平台合作，尤其是在對 AI 技術需求較為迫切的情況下。

不同的 AI 解決方案和工具往往需要與現有的系統協同運作，這使得 AI 開發過程中不可避免地會涉及到複雜性。許多公司通過拆解 AI 系統為模塊化組件來實現簡化，從而降低了實施的難度。

#### (2) AI 與業務流程的融合

將 AI 技術嵌入業務流程本身，而非僅僅將 AI 技術附加到現有流程上，這樣的轉型要求企業在設計工作流和業務模型時，從根本上重新思考如何利用 AI 技術，從而提高效率和創造新的價值。AI 的實施不僅是技術上的變革，更是文化和運營方式的變革，企業需要提供足夠的支持和激勵機制。例如，將 AI 工具納入日常工作流程，並通過績效考核和激勵政策來推動 AI 的普及。

#### (3) AI 生態系統與人才發展

許多專家提到，AI 生態系統的建立應該包括學校、創業公司、大企業和政府等多方力量的協作。企業應該將 AI 技術的引入與創新、商業模式的發展以及人才的培養相結合，這樣才能形成強大的競爭力。對於 AI 人才的需求遠超出單純的技術專業。未來的 AI 專才需要具備跨領域的知識，尤其是在金融、製造和能源等行業，AI 專業人士應該了解業務的深層需求，並能夠將 AI 技術與具體業務流程和目標相結合。

### 四、AI 轉型邁向智能企業

AI 是一項全新的技術，與伺服器、雲端或區塊鏈等傳統技術不同。傳統技術通常解決的是已有問題，而 AI 則是開創全新領域，並要求企業從根本上重新設計其系統和流程。最大問題是從哪裡開始 AI 轉型。大部分人聽說 AI 將來取代工作並改變世界，但具體如何著手，卻沒有明確的指導。講者提供了一個框架，幫助企業設計解決方案和思考未來發展的方向。框架是基於講者的實踐經驗，旨在幫助企業在 AI 轉型的過程中獲得清晰的思路。

AI 的核心挑戰在於，它要求企業徹底重設現有的組織架構、流程和人力資源。現有的系統和流程並不適用於 AI，企業需要基於 AI 的特性，從最基本的原則出發進行再設計。AI 與過去的技术根本不同。它的引入不是僅僅替代某些任務，而是要求一種從根本上改變的方式來構建數據、流程和組織結構。企業需要理解 AI 的本質，並根據其需求重新規

劃組織結構和運營模式。成功的 AI 轉型不僅僅是引入新技術，更是對整個組織進行深層次的變革。

演講者將 AI 比作 19 世紀末的電力革命，並強調了電力引入後，許多工廠仍然選擇傳統的蒸汽驅動系統，因為他們熟悉並依賴這些舊技術。電力發電機組投入使用近 20 年後，大多數工廠仍然使用蒸汽驅動，因為他們對新技術不熟悉，並且過去的運營模式已經穩定。這一點與今天許多企業對 AI 的反應相似，許多企業還未完全理解或準備好迎接 AI。在導入一項新技術時，僅僅更換技術本身是不夠的，整個組織運營和工作模式都需要隨之調整。

將整個工作流程重新設計，是改變效率和提升生產力的關鍵。在電力引入後，工廠能夠靈活地調整工作站和設備的需求，這樣的改變帶來了巨大的效率提升。這一點強調了 AI 轉型與過去技術的不同，AI 並不是簡單地將現有系統更新或升級，而是要求企業從最基本的原則開始重新設計其運營模式和工作流程。

隨著工作流程的分散化（如小型馬達可隨時啟動或關閉），員工的角色和責任也隨之改變。這要求員工具備新的技能，並且更強的責任感。過去依賴傳動系統工廠需要更多技術熟練的員工，而新模式下，員工的職能和工作流程將更加分散，這對人力資源的影響深遠。這樣的變革意味著員工必須重新接受培訓，掌握新的工作技能，以適應更靈活、分散的工作模式。

演講者指出，我們現在正處於類似於“全面電氣化”時代的 AI 轉型階段。正如過去的工廠從蒸汽驅動轉向電力驅動，我們的組織和企業也需要重新設計以適應 AI。演講者提到，未來的方向是人類與機器的協作，就像查爾斯·巴貝奇的差分機一樣，這是計算機和人類合作的早期示例。這一點暗示了 AI 不會完全取代人類，而是成為人類工作的重要夥伴，協助提升工作效率。

就像當初電力革命一樣，AI 的轉型並不是一蹴而就的，企業在轉型的過程中需要重新設計其工作流程、技術架構，甚至是人力資源管理。這是一個深刻且持久的變革過程。演講者通過比喻提醒大家，過去 30 年電力革命中的經驗教訓可以幫助我們理解 AI 轉型中的挑戰——即不僅是簡單的技術更新，而是整體運營模式的重塑。

圍繞著 AI 轉型的實際步驟和策略，並強調了企業面臨的現實挑戰與機會，AI 轉型不可避免會導致一些工作機會的流失，但也會創造新的工作機會。這些變化是無法阻擋的，因此企業必須做好準備。雖然 AI 帶來的變革無法避免，但企業必須清楚哪些領域會變化，以及應該如何改變，以應對這些挑戰。

演講者強調，儘管技術和運營模式會變革，但基礎的需求仍然不會改變——包括對交通、娛樂、通訊、醫療健康等方面的需求。這些需求源自於人類，無論科技如何發展，這些需求始終存在。

變化的是滿足各項需求，需要企業的運營模式進行調整，而企業的基本戰略目標並不會改變，這些策略將幫助企業在 AI 轉型過程中取得成功，**1.**企業的 AI 轉型必須與企業的整體業務戰略相符。**2.**AI 應該用來支持和強化企業的核心目標，而不是偏離企業的基本戰略。**3.**成立一個專門的團隊或委員會，負責 AI 技術的快速整合和實施。這樣的委員會可以幫助協調跨部門的工作，加速轉型進程。**4.**成功的標準不應該僅僅是完成了幾個 AI 項目，而是是否帶來了實際的業務成果。**ROI**（投資回報率）應該是衡量成功的核心指標。**5.**明確定義 AI 轉型對企業的意義，並制定具體的實施步驟和目標。企業必須認識到，轉



型過程中不僅會有成功，還可能會有失敗，因此也需要做好失敗應對的準備。AI 是一個全新的領域，進行 AI 項目就像進行實驗，沒有現成的藍圖可以遵循。企業需要接受可能的失敗並從中學習，這樣才可以不斷迭代並最終找到可行的方案。6.企業應該從最能創造價值的領域開始進行 AI 轉型，而不是嘗試將 AI 應用於所有領域。選擇一個具有規模效應的領域，能讓 AI 技術更快地擴展，帶來可見的價值。7.製定一個具體的計劃清單，列出組織結構中的各個領域，估算每個領域的潛在價值，並進行合理規劃。這是進行 AI 轉型時不可或缺的步驟。過度規劃可能會導致失敗。雖然規劃很重要，但過度的計劃可能會妨礙靈活性，讓企業無法迅速適應變化。因此，在規劃時應保持靈活性和彈性。

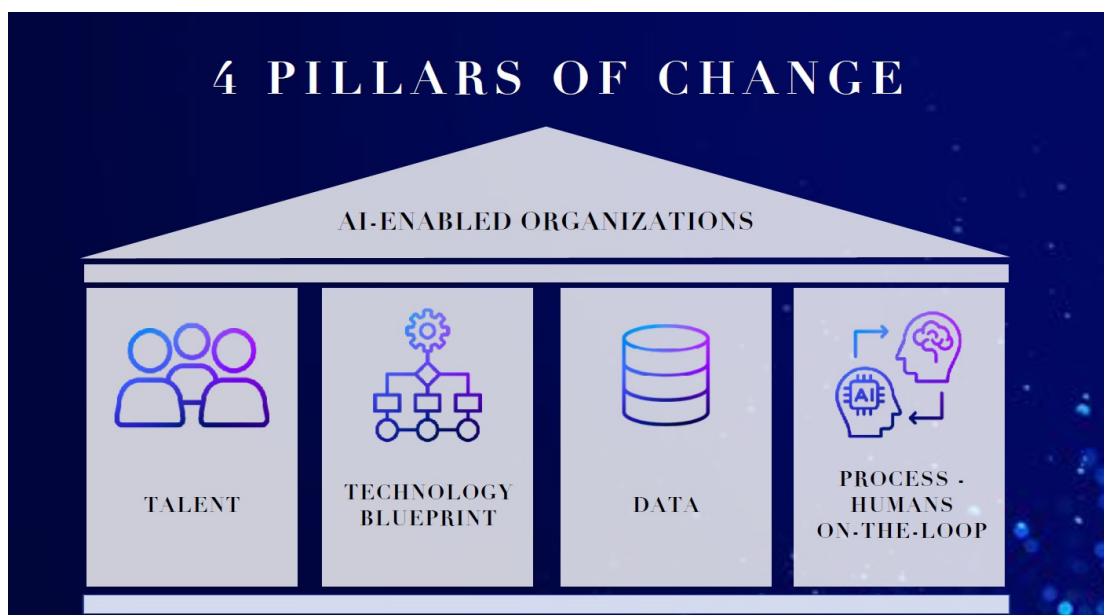


圖四、AI 轉型過程中取得成功的策略

AI 轉型的關鍵步驟和組織必須採取的策略，並提出了實施 AI 技術的具體指導原則，在選擇 AI 轉型的關鍵步驟和組織必須採取的策略，並提出了實施 AI 技術的具體指導原則。

1. 估算價值與可行性篩選 AI 應用領域後，必須進行可行性篩選。這些篩選標準包括高層支持：是否獲得了高層管理的支持？這對於項目的推進至關重要。是否擁有足夠的數據來支持 AI 的開發與運行？開發難度：開發這項技術的難易程度如何？及可擴展性：如果開發成功，這項技術是否具備可擴展性？
2. 評估價值實現的速度，AI 項目應該在 2 至 3 年內實現顯著的商業價值，不能等待 10 年，因為現實中人們的耐心有限。速度應該成為決策的重要考量之一，尤其是在技術轉型中，快速實現成果是關鍵。
3. 評估協同效應，協同效應非常重要，因為它能夠幫助公司基於現有的成果，快速推進後續的 AI 項目。
4. 從小範圍開始，開始時應從小範圍著手，展示 AI 技術的價值，然後再逐步擴大。這有助於降低風險並更快地展示成效。
5. 用運營成果來衡量成功，許多 AI 項目失敗的原因在於它們只關注開發過程，而不考慮是否真正帶來了運營成果。企業應該關注運營層級的 KPI，例如客戶服務效率提升或成本節省等，而不是僅僅完成了 AI 項目。

6. 高層支持與願景的對齊，成功的轉型必須有高層支持。高層應該設定清晰的願景，並將其貫徹到整個企業中。如果沒有高層的支持，轉型計劃將難以實現。與高層的願景對齊，並確保持續的內部溝通，對於 AI 轉型至關重要。
7. 設計解決方案的原則，設計解決方案時，應遵循以下四個關鍵原則：
  - 速度：AI 可以極大提升速度，並且可以幫助企業更快適應市場變化。
  - 適應性：AI 支持的組織應具備快速適應市場變化的能力。
  - 規模：AI 技術具有規模化的潛力，一旦成功，應該可以迅速擴展。
  - 實驗：AI 轉型是一個持續實驗的過程。企業應該不斷進行小範圍的測試，找出可行的方案，並推向全公司。



圖五、組織變革的 4 大支柱

AI 應該能夠幫助組織在變化的環境中自動調整，變革包含：

人才：企業轉型的核心在於人才。AI 將取代一些重複性工作，但仍需要人才來驅動創新和處理複雜的問題。

技術：基礎設施必須支持快速適應、快速擴展和持續實驗的能力。解耦架構、自主性和自動化是 AI 技術基礎設施的核心特徵。

流程：數據系統和流程的設計至關重要。AI 需要一個精心設計的數據結構來進行有效的學習與預測。

企業的目的是滿足客戶需求，而非僅僅創造就業機會。AI 的核心在於如何用它來提升服務質量、加速決策和提升效率。AI 轉型的多個方面，包括評估 AI 項目價值的標準（如高層支持、數據可用性、可擴展性），並提出了幾個關鍵原則來設計 AI 解決方案，如速度、適應性、規模與實驗。強調了高層支持的重要性，以及 AI 技術對組織流程、人才和技術基礎設施的深刻影響。最終，企業需要利用 AI 來快速適應市場變化，並創造實際的商業價值。

## 五、數位平台公司轉型(AI 代理)

在學習和使用 AI 之前，選擇一個可靠的、運行良好的平台至關重要。這不僅關乎供應商和提供方的選擇，還需要在工具和資金等方面達到良好的平衡。OnePage 平台最初由創

業者引入，並設計為一個推薦系統，通過高質量的網絡推薦來提升效率。這樣的推薦系統能夠通過 AI 模型提高預測精度，並實現自動化運作。

AI 正在使用生成型 AI 技術來進行轉換和優化，這不僅是簡單的數據處理，還包括更複雜的定位和預測。這樣的技術使得系統能夠更精確地進行「切割」，即對數據和服務進行細化處理。

公司通過穩定的擴散技術快速開發虛擬角色原型，並與 GPT 和 TTS 技術相結合。這種快速原型開發的方式使得服務能夠在短短一個月內完成並投入使用。

推出了名為 InterviewView 的虛擬 AI 面試應用，這個應用不僅能幫助用戶優化簡歷，還能進行自動翻譯和職位推薦，進一步改善求職者的工作準備過程。公司正在開發職業地圖，幫助用戶規劃職業路徑，並根據用戶的職業發展需求提供相應的指導。提供如何填寫職位申請表的指導，並根據簡歷和職位要求推薦合適的工作，這種服務使求職者能夠更有效地應對求職過程。

AI 的開發過程與傳統 IT 開發過程存在本質區別。在開發 AI 服務時，傳統的功能性環境和非功能性環境的區分不再適用，這要求企業對開發流程進行調整和創新。服務質量與專業性：服務的質量需要依賴領域專業知識，並且 AI 需要具備足夠的專業性來解決特定領域的問題。AI 生成的內容和模型可能存在 10% 的“幻覺”問題，即出現不準確或不一致的情況，因此需要不斷的測試與改進。

雖然公共 NLM（大型語言模型）並不是最強的，但其綜合性和增強的安全性讓它成為企業選擇的理想方案。隨著公共雲服務的快速發展，企業逐漸轉向使用公共雲來支持其 AI 和數據處理需求。選擇合適的開發環境和模型對於企業至關重要，因為這會影響到成本、服務質量以及未來擴展的可行性。許多公司專注於保障 AI 服務的安全性，特別是涉及到個人數據時。對於 AI 服務的安全性，許多企業都已經投入了大量的資源來強化這方面的防護。

在整合 AI 和傳統 IT 系統時，會遇到許多挑戰，尤其是在服務質量和運行效率之間的平衡。AI 的生成與處理結果需要進行持續調整，以保持系統的穩定性和服務的專業性。隨著 AI 技術的進一步進化，企業應該關注如何使用 AI 來解決實際問題，並尋找最佳的模型和開發環境來支持其業務發展。網站文檔的創建需要關注多個層面的內容，包括模型的選擇以及未來如何根據需求更新或改進。關鍵是要有清晰的優先級，這些文檔和模型會指導產品的開發和改進方向。產品類型：需要確定所涉及的产品是否為示範產品或附加產品，並了解它們是否通過多終端渠道來分發。

為了讓測試過程更高效，必須使每個測試步驟都能輕鬆操作，並且每次測試結果都能提供有價值的反饋。透過預期結果的比較，快速作出決策，這樣能更好地控制產品開發的方向和質量。目前大多數平台的搜索結果是基於字典或預定規則生成的，但缺乏專家知識或數據支持，這可能會影響結果的準確性。

平台應該支持協作，並且能夠讓開發者和前端工程師之間實現無縫共享。即使開發的是一次性的產品，也能夠確保在市場變動時不會過時。創建共享的工作環境，這樣不同部門可以進行無縫協作。這種環境不僅限於合作夥伴，還可以涵蓋內部不同團隊的協作與信息流通。

為了能夠跨多終端平台順利運行，需要開發一個單一的 LLM（大型語言模型），該模型能夠支持所有功能並高效運行。此外，這個系統應該能夠進行強大的後端支持，如監控、

維護等。特定的功能平台，可以讓公司輕鬆進行產品的編輯和開發。這樣的系統使得不同區域的功能和資源能夠高效共享。

**快速開發與迭代：**透過平台的支持，開發團隊能夠在較短時間內（例如 3 天）完成一個服務或產品的開發並進行測試。這樣的迭代速度是當今快速變化的市場環境中非常重要的一個特徵。內部開發的 AI 平台，用於支持產品開發和測試。它與 Google Workspace、GitHub 等工具集成，有助於協作和產品性能測量。

公司在轉型過程中進行了內部審計，確保不同部門能夠自然地進行協作，並將這些轉型過程順利地交付到所需的系統中。新的系統和流程設計旨在保證不會影響到已有員工的工作模式，從而實現順利的過渡。

公司已經開發了 30 個 AI 機器人和 20,000 個 SQL 生成器，並且利用這些工具來進行日常的數據生成和管理。這些生成工具對於快速數據處理和決策支持至關重要。SQL 生成器被運行了超過一年，並且是公司內部數據操作的核心工具之一，顯示出 AI 在數據管理領域的巨大價值。

公司在過去的十個半月內取得了顯著的成長，並成功推出了多個關鍵產品。隨著 AI 和其他新技術的融合，許多舊有的流程和技術正在被逐步取代，以支持更高效的運營模式。網站和平台的文檔創建過程中，選擇合適的模型、產品類型和測試流程至關重要。通過優化這些流程，公司能夠快速開發和測試新的產品。

AI 技術在提升產品開發效率、數據生成和管理方面發揮了關鍵作用，並且公司正在不斷進行迭代，提升這些技術的應用範圍。創建能夠支持不同部門協作的共享工作環境，從而提高開發和測試的效率，並促進產品的快速迭代。在快速開發的背景下，測試和質量管理流程依然是關鍵，通過精細化的 QA 檢查，公司確保了產品的穩定性和高效性。

**創意競賽與團隊協作：**公司進行了多個團隊的創意競賽，以推動業務創新。各團隊的合作與競爭激發了創新的想法，並且使得他們能夠快速應對過程中的挑戰。儘管公司仍在努力提升業務運營，但過去的經驗幫助他們在面對挑戰時更加靈活，且能夠持續改進現有流程。

開發一個能夠高效支持各類用戶的系統，這不僅是為了平台的準備工作，也是為了未來的業務增長。雖然這一過程中充滿了試錯，但能夠快速學習和調整是關鍵。公司在平台開發上投入了大量的精力，並且認識到這些平台將支持未來的多樣化業務需求。

**AI 代理的概念與應用：**

AI 代理是一種能夠自動執行任務的系統，根據不同的應用領域，它需要具備足夠的理解能力和靈活性。這些代理可以被設計為執行簡單或複雜的操作，並且能夠通過提示來控制其行為。AI 代理的發展分為幾個階段，從基礎的自然語言處理（NL）到更加複雜的產品應用。第一階段主要是基於市場數據的初步應用，而更高級的代理系統將會在多個領域進行協同工作。平台需要有能力支持和協調多個代理，使它們能夠高效協同工作。這不僅是技術的挑戰，也需要更高層次的協同管理來確保運行的順利。

**AI 代理的管理與挑戰：**在管理 AI 代理時，公司面臨著許多挑戰，尤其是自動化過程中的“黑箱”問題。自動化雖然提升了效率，但也可能帶來無法完全理解的結果。因此，需要小心管理這些代理的運行，並且不斷進行全面檢查。儘管自動化是一個強大的工具，但完全自動化也可能消除一些人類操作的靈活性。過去，公司曾經避免過度自動化，但隨

著行業的發展，許多自動化流程變得不可避免。這要求公司在自動化與人力操作之間找到平衡。

**AI 平台的基礎與挑戰：**AI 平台的基本原則是能夠檢測並反應各種信號，這些平台不僅能創造內容，還能夠基於實時數據進行決策和調整。這樣的平台可以進一步促進創作過程的擴展和深度發展。儘管自動化是提升效率的重要手段，但過度依賴自動化可能會帶來風險。黑箱問題使得決策過程不夠透明，且無法確保所有操作都能符合預期。

公司在過程中進行了大量的測試，並且通過不斷的嘗試和調整來尋找最佳解決方案。這種“試錯”的方式雖然充滿挑戰，但也促進了創新和改進。公司計劃在未來六個月內舉辦見面活動，以便展示其在 AI 平台和自動化領域的最新進展。這也是一次展示成就和了解未來方向的機會。

未來市場上將出現多種類型的 AI 代理，這些代理將根據各自服務的目的和領域有所不同。每個代理的設計和應用將密切與其所服務的需求對接。為了實現最佳的運行效果，AI 代理需要具備良好的協調機制，這樣才能確保在實際應用中達到高效運作。

## 六、AI 應用領域

### 1、生成式 AI 與搜尋介面的評估—以 Naver 系統為例

Naver 的搜尋介面經歷了多次變革，從最初的搜尋框，到綠色框的品牌化，再到現在基於 UX 和 AI 技術的創新。Naver 不斷更新其搜尋框，搜尋歷程從最初的網頁目錄到現今更為精確的搜尋引擎，展示了搜尋體驗如何隨著技術進步而進化，越來越強調與用戶需求的對接。強調科技如何影響品牌形象和用戶體驗，並且指出每次科技突破都可能會帶來全新的設計方向。

當機器學習逐步進行商業化，未來的應用範圍會有多大。Naver 團隊通過與 AI 技術的結合，正在探索不斷擴展的應用場景，尤其是在 UX 設計的層面上。如何將 AI、UX 設計以及搜索技術結合，打造出更加精確、具有品牌特徵的產品界面，並利用這些創新提升用戶體驗。

Naver 的服務於 2009 年開始支援行動端，改變了使用者的互動方式。介面不再是單純的小型資訊介面，而是轉向能夠在行動裝置上觸控操作、適應手指滑動的版本。行動端的普及使得 Naver 能夠收集更多數據，並根據這些數據了解用戶需求，從而更精確地調整搜尋結果。Naver 開始根據用戶需求對數據進行分類，例如針對汽車、旅遊、遊戲、電影等主題提供更貼近使用者的搜尋介面。例如，針對電影數據，Naver 不僅展示電影本身的資訊，還能將其與導演、動畫、音樂等相關資料聯繫，讓使用者能更全面地了解相關內容，從文字到圖像，從圖像到影片，現在則出現了短影音形式，並且 XR（擴增實境）等新型內容也開始被引入到搜尋服務中，不再僅僅顯示文字或圖片，而是能以即時互動的形式展現暴龍的動作與姿態，為使用者提供身臨其境的體驗。

隨著人工智慧技術的發展，Naver 能夠根據使用者的需求和行為進行個性化推薦，提供更符合個人偏好的搜尋結果。隨著 AI 的進步，搜尋結果會更精準地根據他們的需求、偏好和使用行為進行差異化推薦。隨著 AI 技術的應用，Naver 的搜尋結果開始變得更加個性化。舉例來說，對於兩位不同的使用者，即使搜尋同一個物品（例如冰箱），系統會根據他們的搜尋歷史與需求提供不同的結果。這使得搜尋結果更加精確、符合用戶的實際需求，並且持續更新以適應不斷變化的需求。Naver 推出了基於 AI 的「Q for AI」搜



尋引擎，這是一個新一代的搜尋工具，透過與 AI 的互動來提供搜尋結果。這不再是傳統的關鍵字搜尋，而是支持用戶輸入更長的句子，進行更精準的搜尋。搜尋不再僅依賴簡單的關鍵字，而是變成了對話式的搜尋體驗，使得結果更符合用戶的需求。

隨著時間的推移，搜尋問題的提問方式逐漸變長，這與用戶在線上交流中使用「聊天語言」的習慣有關。這一變化使得搜尋引擎需要適應並改變其對話設計，從而讓使用者更容易理解搜尋結果。儘管對話式搜尋逐漸流行，但對話次數減少的趨勢表明，搜尋行為仍在不斷發展，這也意味著搜尋服務需要繼續調整其功能以適應新的用戶行為。

隨著 AI 生成內容的普及，Naver 強化了資訊來源的標示，確保用戶能夠了解自己所查看內容的來源及其可靠性。這一點對於提升資訊的可信度至關重要，並且在 UI/UX 設計中進行了加強，讓用戶清楚地知道他們所接收的信息是來自可信的來源。

Naver 將搜尋結果與其他服務（如購物、預約等）進行更深層次的整合。例如，當用戶搜尋與 CrossFit 相關的內容時，可以看到推薦的產品（如三星智慧手錶），並且可以直接購買或查看相關應用程式，提升了搜尋引擎的服務深度。這樣的整合使得用戶不僅能找到資訊，還能直接完成購物或預約，極大地提高了搜尋的便捷性。

Naver 正在改進其設計和開發流程，將設計資源（如範本和元件）進行代幣化，並建立了與開發流程緊密相連的系統，旨在提高設計師和開發者的生產力。這種新的設計系統（設計系統 2.0）將使得設計過程更加高效，並支持更多創新。

隨著 AI、LLM、大型語言模型等技術的發展，Naver 的搜尋引擎和服務將繼續演進，並且將不僅僅限於搜尋資訊，而是將涵蓋購物、預約、娛樂等多方面的功能，形成一個更加綜合的服務模型。Naver 還將不斷探索新的服務形式，如車載系統、智慧音箱等，並將這些服務與搜索引擎進行整合，實現更大的協同效應。

## 2、視覺創作

AI 在視覺創作中的影響，改變了視覺藝術和創作過程，AI 也在改變影片製作的方式，特別是在故事講述和創意表達上。AI 技術不僅僅是在創建視頻或圖像，而是在推動更深層次的創意探索，提供更多的控制權和靈活性。

Runway(講者)專注於開發通用 AI 模型，這些模型能夠理解和推理世界及其動態，並不僅限於視頻生成。這些模型還涉及音頻、文字、圖像等不同形式的數據，並能模擬世界的各種情境。目的是建立一個強大的模擬系統，為不同創作領域提供支持，無論是電影製作、廣告、還是其他視覺創意工作。

AI 技術能夠在電影和視頻創作中創建那些以前無法實現的場景，特別是涉及過渡複雜和視覺效果的場景。例如，從叢林過渡到圖書館，再到牛群和大海的場景，這樣的效果通過 AI 模型能夠平滑且高效地呈現。AI 技術的發展不僅限於渲染超現實視頻，還能夠結合各種元素，如動畫和實景拍攝，創造出更多元的視覺效果。這不僅是為了提高創作效率，更是為了激發新的創意和視覺語言，打破傳統創作的界限。這些 AI 工具對藝術家來說，不僅是一種技術支持，還能成為創作過程中不可或缺的合作夥伴。

AI 模型強調藝術性和創意控制，這對電影製作人和其他創作者至關重要。儘管 AI 生成技術強大，但藝術指導的角色仍然是必不可少的，確保 AI 生成的內容符合創作者的情感和美學需求。生成式 AI 在電影和視覺創作中的應用，正為故事講述提供新的方式。AI 不僅能創建具體的視覺效果，還能在情感層面與觀眾建立連結，使故事表達更加多樣化和動人。



總結來說，致力於將 AI 視覺創作提升到一個新的高度，這不僅僅是技術的進步，更是創意和故事講述方式的一次革命。隨著 AI 的發展，電影製作和其他視覺藝術領域將迎來更多創新的可能性，從而改變藝術家和創作者的工作方式，並為觀眾呈現出更多元、深刻的視覺體驗。

AI 技術的發展使得創作過程變得像魔法一樣。從傳統的提示語言（如文字描述）到生成複雜的視頻和視覺效果，這種轉變使得創作者能夠探索過去因技術限制而無法實現的創意方向。AI 能夠迅速渲染過渡效果，為創意探索帶來了新的機會。這促成了當代的創意探索新文藝復興，藝術家和創作者不斷進行實驗，發現全新的創作方式。

AI 生成的視覺效果不僅能夠精確控制，還能在創作過程中進行實時修改和調整。例如，生成的角色動畫可以進行調整，並保持與原視頻的一致性，創作者能夠快速改變某些部分，如背景、動作、甚至劇情發展。這種即時修改和高控制性使得 AI 技術成為非常強大的創作工具，能夠有效支持視覺特效的創作，並且在電影和遊戲等領域提供無限的創意可能性。

Runway 的技術已經開始與好萊塢工作室、流媒體公司和其他創意產業的領頭羊合作，通過使用數據來創建虛擬世界，並用這些世界創建更多內容。這些 AI 模型支持快速的實時定制，能夠在電影製作過程中進行即時的視覺特效渲染和場景變化。例如，在視頻生成過程中，能夠實時將現場動作與生成的視頻特效結合，從而實現更加靈活和高效的創作過程。

AI 技術的發展為視覺創作帶來了新的工具和可能性。就像過去相機的發明改變了電影和故事講述一樣，AI 將繼續推動創意的邊界，使得故事講述和視覺藝術的創作變得更加多元化和高效。隨著技術的進步，AI 不僅是創作者的工具，它將成為創作過程的一部分，

### 3、人形機器人

機器人技術的歷史，從 60 年前的 Unimate 機器人開始，這是一個液壓機械臂，能夠執行簡單的物品搬運任務。隨著時間的推移，從液壓技術到電動技術的轉變讓機器人擁有了更多的靈活性和精度。特別是在自動化技術和視覺識別的應用上。隨著無軌機器人（AMR）和自主移動機器人（AGV）的出現，機器人開始在物流和倉儲設施中發揮關鍵作用。例如，亞馬遜使用超過 75 萬台這樣的機器人來自動化其倉庫操作，實現了無人搬運和自動化管理。但許多工作仍需要人類來執行，特別是那些髒、重、危險且重複的工作。目前美國有超過 100 萬個工作崗位空缺，主要是倉庫和製造業的低階工作，這對企業造成了極大的挑戰。

隨著對電商隔日送達需求的增長，傳統的人力資源已無法跟上需求的增長速度。因此，人形機器人被視為解決方案，人形機器人不僅限於工業生產，它還能夠進入家庭領域。未來，消費者可以將機器人派去做一些他們不願意或無法做的家務活，如修剪草坪、洗衣等，這將改變人們的生活方式。這項技術已經開始進行商業化，"機器人即服務"（Robot-as-a-Service）已經開始在某些領域落地。

機器人不僅僅是獨立作業，它們正在協同工作，兩台機器人間能夠連續運行並協作完成某些特定的任務，如物品搬運、包裝等。在這些合作中，機器人需要具備高精度的感測和動作控制技術，使它們能夠處理各種複雜的物理任務。

人形機器人如 **Digit** 通過人工智慧，能理解語音指令並能夠執行複雜的任務，例如掃描環境，避免碰撞，並在家庭或企業環境中執行簡單或複雜的操作。目前，人形機器人主要集中在企業市場，尤其是在倉儲、製造等高需求領域。這些機器人需要能夠理解指令並與人類協同工作，特別是在安全性和高效性的平衡上。語意智能使機器人能理解指令，並執行相應動作；物理智能則指機器人如何動作、如何利用機械部件（如手臂、腿部等）完成任務。

**Digit** 成功執行了分開白色與有色衣物的任務，展示了它在家庭環境中的潛力。儘管目前主要針對企業市場，但未來幾年內，這類機器人可能會進入家庭市場。隨著技術的成熟和成本的下降，人形機器人有潛力進入家庭市場，並承擔家務工作，如做飯、洗衣等。希望有一天機器人能在家中幫忙做飯和家務，這樣可以大大減少家務勞動的時間，提升生活質量。

隨著協同安全的技術進步，**Digit** 等機器人能夠與人類在同一環境中協作，進行更精密的工作。協同安全是當前人形機器人的最大挑戰。機器人必須確保在與人類一起工作的環境中，不會對人類或周圍環境造成傷害。機器人目前的電池續航能力是另一大挑戰，但隨著新型電池技術（如基於石墨烯的電池）的引入，未來的機器人將具備更長的工作時間和更快的充電速度。人形機器人主要目的是減少人類工人負擔，尤其是那些危險、單調或重複的工作任務。這類工作常常導致工人受傷或疲勞。機器人的引入能幫助企業減少人力資源的需求，並降低工作中的風險，從而提升整體生產力。

工作機會的減少是熱議話題。佩吉強調，機器人不會完全取代人類工人，而是幫助減輕一些工人不喜歡的工作部分，尤其是危險、疲勞或單調的工作，並且使工人能夠集中精力從事更有創造性和價值的工作。人形機器人技術的快速進步，尤其是在人工智慧的協助下，這些機器人不僅能夠執行更多的任務，而且能夠與人類更安全、更高效地合作。隨著技術的進步，未來幾年，這些機器人可能會進一步進入家庭市場，並成為日常生活的一部分。

#### 4. LG AI 代理系統設計與發展

討論 AI 代理的設計與開發過程，並解釋它們如何在電子產品中發揮作用。AI 代理 被定義為具有自主性、能夠識別環境並根據情境做出決策的系統。其目標是解決實際問題。類似人類解決問題的過程：識別問題、收集信息、選擇最佳解決方案。在 AI 代理設計中，則是結合 多模態 的感知方式來更全面地理解環境。使用語音、視頻、聲音感測器等來識別情境，並根據不同的模態來分析問題。識別問題、收集信息、選擇最合適的解決方法。AI 代理需要處理這些步驟，並且從 多模態 的信息中提取決策。

LG 對 AI 代理的關注特別集中於 家庭 生活領域，目的是提升家庭智能家居的價值。目標是開發一個能夠超越現有智能手機的 高端 AI 家庭助手。同理心的 AI 代理：與目前大多數基於功能的智能家居助手不同，LG 正在開發具有同理心的 AI 代理，作為家庭中的「夥伴」和「陪伴者」。這些 AI 代理不僅執行基本的家居功能，還能夠處理家庭中的各種問題，並能夠與使用者進行深層的情感互動。在 MBTI 人格類型中，LG 的 AI 代理更像是 F 型（情感型），而不像傳統的 T 型（思考型），強調情感和共情能力。

現有 AI 代理 大多僅提供簡單的設備連接功能，但 LG 想要的是一個能夠深度理解家庭環境並解決實際問題的代理。每個家庭部門（例如廚房、客廳等）都有需要解決的問題，LG 希望透過智能代理來協同解決這些問題，並提升生活品質。

通過多種感測器來接收來自不同設備的 多模態感測數據，並將這些數據用於理解當前情境。系統會使用模型來理解當前的情境，並根據理解來選擇相應的工具和 API。系統的核心是根據 情境理解 選擇最合適的反應方式，這能夠讓 AI 代理更加靈活且精確地應對多變的家庭需求。

LG 的 AI 代理系統不是單純的功能拆解，而是通過 化學整合（指多個模組間的緊密協作）來實現高效的功能和活動。不同部分（例如智能感知、情境理解、反應協同）協同工作，實現整體高效的動作。系統設計將著重於簡單、直觀的用戶界面，讓用戶能夠清楚理解代理所選擇的內容，並能夠以自然語言進行互動。

LG 的 AI 代理系統不僅致力於提升家庭智能化，還將強調情感、同理心和多模態感知的整合。這些 AI 代理將成為家庭生活中的真實夥伴，從簡單的設備控制到深層的情感交流，逐步解決家庭中日常生活的複雜問題。LG 的最終目標是創造一個能夠真正理解並反應家庭情境的系統，並提供無縫的 現實世界 和 數位世界 的連接體驗。

**LG 多模態 Alzert 系統介紹：語音與智能感知應用**

LG 在語音與多模態 AI 代理 領域的發展，特別是將如何將 語音識別、情境理解、智能感知技術 等技術應用到 智能家居與健康安全中。QNN Alzert 是一個多模態的 AI 系統，將語音、視覺、感測器數據等結合起來，提升家居環境的智能化。系統能夠根據用戶的需求和情況，調整家居環境。例如，在學習時，系統根據氛圍調整燈光、溫度、空氣濕度等參數，創造適宜的學習環境。

不僅支持語音和對話，還能根據用戶的行為（如外出、咳嗽、空氣質量差等）調整家居環境。例如，當發現空氣質量差時，代理人會啟動通風系統或更換空氣過濾器，限制病毒傳播，改善家庭健康。

關鍵在於收集家庭空間中的各種數據，這些數據被用於環境理解和自動調節。LG 使用的感測器包括 阻抗傳感器、空間傳感器、化學傳感器 和 毫米波傳感器。例如，空調中的毫米波傳感器可以感知用戶在房間中的位置和動態，從而調整空調的運行方式。

這些傳感器能檢測空氣中的有害物質（如甲醛等），並根據檢測結果作出對應的反應。利用視覺技術識別用戶的臉部狀況、專注度等，並調整相應的家居設備，如烤箱、冰箱等，進行自動調整。

基於先前收集的數據，AI 系統會進行情境分析，理解用戶的需求並調整環境。例如，當用戶在家時，語音識別技術可以讓系統理解他們的需求並推薦適合的內容（如電視節目等）。LG 特別關注如何利用這些技術來增強家庭安全和健康，這是未來發展的一個重要方向。

LG 正在深入研究語音轉文字、語音合成和文本轉語音的表現，這將使語音交互更流暢自然。強調語音處理的整體流程，確保語音識別和生成的高效性與準確性。這是一種處理指令和執行設備反應的技術，通過這種技術，只有當設備能夠執行特定用戶需求時，它才會作出反應。LG 利用 LEG 技術 來提升語音識別的延遲和準確性，並將這些技術應用於 產品 QA（例如，故障診斷、手冊指導等），提升用戶體驗。

在進行智能協同操作時，系統需要選擇最合適的工具並進行反應。在這個過程中，準確度、延遲和成本 是主要的衡量指標。例如，對於一些高成本的設備來說，延遲可能

非常昂貴，而低延遲的回應能提高整體效率。低成本與高效能模型：LG 正在使用類似 GPT 這樣的大型模型來提供強大的運算能力，同時也關注低成本的模型來平衡效率與成本。

AI 插入技術：LG 的最終目標是實現 AI 插入、設備與服務的整合，使其成為智能家居的核心。開放的測試服務：今年，LG 將提供測試服務並開放部分關鍵字，讓用戶在日常使用中檢查和反饋 AI 系統的表現。LG 的多模態 AI 代理正在實現一個能夠感知、理解和協同工作的智能系統，並將這些技術應用於家庭生活中，提升用戶體驗。特別是語音識別、情境理解、智能感知等技術，將大大增強家居設備的智能化，並實現健康、安全等領域的應用。LG 的目標是實現無縫連接的智能家庭，並為用戶提供全方位的服務和支持。

## 七、硬體設備(半導體)需求

強調沒有半導體，AI 無法實現，半導體是支撐 AI 運行的基礎硬體。這一點顯示了 AI 和硬體（特別是半導體）之間的密切關係。這可能在強調 AI 的高效能與強大處理能力。這或許是在描述 AI 可以達到的無所不知與無所不能的潛力。轉換模型(transformer model) 被認為是目前 AI 領域最重要的模型之一。強調編碼器和解碼器的角色，並提到這些模型如何處理和生成信息，並且指出在未來 30 年內，這一模型應該不會發生大變化。

提到將 GPU 放置在家中的重要性，並指出 GPU 在當今 AI 研究中扮演的核心角色。這也體現了 AI 研究者對硬體（尤其是 GPU）的依賴。主講者談到 AI 在音樂、文字、電影等創意產業的應用，並且提到未來的影片創作將不再依賴人類，而是由 AI 來完成。引用了 Sam Altman 和 Google CEO 的例子，暗示未來的影片創作和信息流通會由 AI 來主導，這可能會徹底改變 YouTube 等平台上的內容創作和消費模式。

運用 AI 來自動化 HBM（高頻寬記憶體）的設計，並展示了 AI 在這一領域的潛力。這部分突顯了 AI 在硬體設計領域的應用，尤其是提升設計效率和準確度。AI 不斷進步，但目前還有很多挑戰，尤其是計算資源的需求。AI 需要大量的記憶體、處理能力和電力來進行訓練和運行。AI 技術的核心在於處理和存儲大量數據，尤其是在深度學習中，GPU 和記憶體（如 HBM）的發展至關重要。記憶體管理將決定 AI 系統的性能。半導體產業對 AI 技術至關重要，尤其是記憶體技術（如 HBM、SDM、DDR 等），它們決定了 AI 處理的速度與效能。目前，GPU 與記憶體之間的配合有很多挑戰，尤其是在存儲和計算效率方面的平衡。記憶體的發展趨勢是將更多功能移至記憶體上，以提升計算效率並減少延遲。隨著 AI 技術的進步，未來可能會出現一個新的經濟模型，其中 AI 與人類共存並協作。

AI 將對許多行業（如電影製作、設計等）產生深遠影響，但同時也會帶來需要大量資金支持的挑戰。金融支持將是未來 AI 發展的重要議題，尤其是數據中心和能源消耗方面。未來的記憶體技術（如 HBM）將對計算過程至關重要，並有可能改變 GPU 的設計與使用方式。隨著計算需求的增長，如何處理 AI 硬體設備的熱量成為一大挑戰。液冷系統可能成為未來數據中心的一個解決方案。

AI 的自動化和大規模服務將提升效率，但如何保護人的情感、倫理和意識將是未來的一大挑戰。AI 可能會對工作場所、經濟模式以及社會結構產生重大影響。半導體、記憶體和 GPU 是 AI 運行的基石，對這些技術的投資將決定未來 AI 的發展。AI 將改變很多領域，但人類的創造力和情感無可取代，未來的挑戰是如何平衡 AI 和人類的角色。大量的記憶體、運算能力以及能源將是 AI 發展的最大挑戰，需要巨大的投資來支持這一變革。

## 八、語言模型

### 1、LLM 模型的測試與評量

AI 隨著計算能力的提升，語言模型（如 GPT 系列）變得越來越強大。從 2018 年到現在，這些技術的發展速度驚人，模型的參數規模和性能持續增長。以 GPT-3 和 GPT-4 為例，這些模型的參數規模超過了數兆，推動了更多智能產品的誕生。

「推理原生模型」的出現，這是突破了傳統 IQ 測試的方式，能夠在多層次的問題解決中超越人類平均水平。例如，OpenAI 的 O1 模型成功在 IQ 測試中達到 120 分，超越了先前的模型。透過排行榜或測評工具（如 UC 伯克利創建的棋盤格），可以比較不同語言模型的性能。這些測評基於模型的得分，會隨著技術的進步而變化。

模型效能在不斷提升，成本卻會隨著技術進步而下降。這要求用戶和企業不斷關注性價比，選擇合適的模型平台，因此在選擇模型時，需要仔細計算每次運行的 token 數量以及相應的費用。不同模型的價格差異可能非常大，根據模型的效能和 token 數量來選擇合適的解決方案。即使性能更好的模型價格較高，但如果能顯著提高效率，則可能是值得的。

AI 模型的使用不僅限於學術研究，還在商業中展現了強大的潛力。例如，AI 的知識庫管理、數據摘要、自然語言處理等功能已經被廣泛應用。隨著新技術和新模型的推出，市場的競爭變得更加激烈。這也意味著不斷測試和改進自己的系統，以應對不斷變化的需求和挑戰。

過去公司可能只依賴一兩個大模型（如 GPT 或亞馬遜 AWS）。但如今，使用多個模型變得越來越普遍。許多開發者會選擇 3 個、5 個甚至更多的模型來完成項目。開發者現在更依賴 AI 來撰寫代碼，並且結合公司的知識庫來開發 AI 服務。提示引擎仍然是不可忽視的工具，儘管 AI 技術有了更多進步，但理解和運用提示引擎仍然有其重要性。企業可以將公司內部的知識資料轉換成向量空間，通過自然語言嵌入 API 來進行檢索。當問題轉換為向量後，語言模型就能根據這些向量找到相關資料來回答問題。

使用圖譜分析工具來建立公司文檔的結構化知識圖譜，這樣能有效地分析和提取資料，解決更復雜的問題。例如微軟推出了開放的語義核心工具，這些工具用來處理語義記憶。這些工具支持更本地化的資料處理方式，並能提升性能，不再完全依賴 Python，也使用 C# 來加速處理。隨著 AI 技術的發展，現在許多模型不僅處理文本，還能處理圖像、語音、視頻等多種數據形式。這些所謂的「視覺影像模型」具備多模態能力，能同時處理圖片和文本，並且能進行有效的識別與推理。

系統中代理扮演著重要角色。每個代理都有特定的職責，並且能夠互相協作來達成目標。這些代理可以是專家代理（如資料科學家、技術架構師、事實檢查員等），他們的協作幫助完成複雜任務。多代理系統的協調非常重要。系統會有一個「任務管理器」，負責分配任務並監控每個子任務的進度。這是一種高度協作的工作方式，讓代理能夠在分工的同時維持高效的協同。

這樣的多代理系統不僅能解決特定的技術問題，還可以處理複雜的決策過程。代理不僅是獨立的工作單位，還能互相配合進行多層次的協作。“定制代理”強調的是根據特定需求創建的代理系統。用戶可以針對具體場景創建自己的代理，並調整它們的功能以滿足需求。這種方式為用戶提供了靈活性和可擴展性。

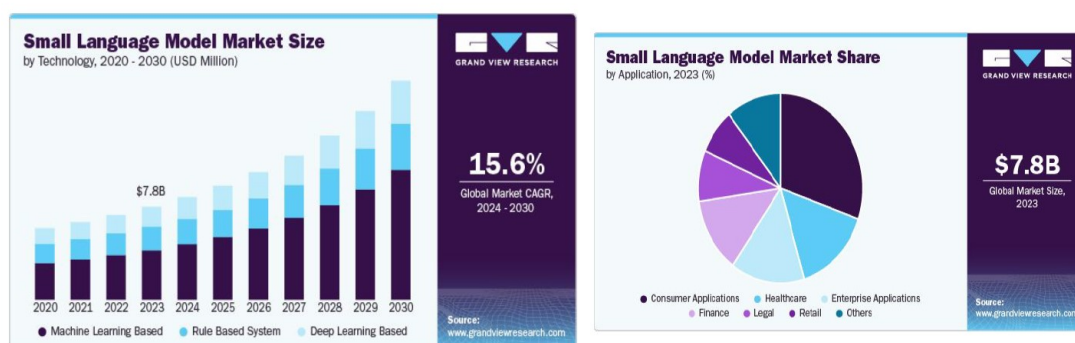
Copilot Studio 是微軟提供的一個工具，允許開發者創建和定制自己的代理。這些代理可以像創建自定義 GPT 模型一樣被設計並部署，並且能夠與現有系統（如 SharePoint、Dynamics 365）協同工作。當用戶需要處理現場服務或服務呼叫時，這些代理可以與動態 365 的後端系統進行集成，自動處理工作流和事件生成。通過多代理協作，工作可以以更高的速度和效率完成。這不僅提高了任務完成的質量，還能加速決策過程。

基於量化方法提升運算效能、以及如何利用 代理系統 和 多代理協作 來高效完成複雜任務的理念。這些代理不僅在處理單一任務時能協作，還能進行靈活的定制，並且結合現有的 IT 系統（如 Dynamics 365、SharePoint）實現完整的工作流自動化。最終的目標是通過這些技術達到更高效的結果，並且讓人類能夠專注於更有價值的工作，從而提升整體的生產力。

## 2、小型語言模型與神經網絡

對於小型神經網絡模型（SNM），市場和學術界的定義仍然不一致，這是因為不同的機構使用不同的標準來評估模型的大小和功能。小型語言模型相比大型模型，在隱私敏感領域（如國防、法律、醫療等）有明顯的優勢。這些模型可以在本地部署，避免將數據傳送到外部伺服器，從而確保數據安全。此外，小型模型的運行成本較低，對硬體要求也較小，這使得它們更具可操作性。模型之間的區別不僅僅在於它們的大小，而是取決於它們是如何訓練的，或者它們的使用目的。

### SLM market is expected to be \$20B market by 2030



圖六、小型語言模型市場增幅

小型語言模型的需求正在快速增長，並且越來越多的大型科技公司推出了自己的小型語言模型版本。未來幾年，這個市場將進一步擴大，預計將達到 80 億美元，並且將增長至 200 億美元。儘管小型語言模型在效率上有優勢，但它們仍然需要大量的能源來訓練。這些訓練過程消耗了大量的水和能源，這對環境造成了巨大壓力，這讓人對人工智慧的發展產生了深刻的擔憂。對於能源問題，專家們認為，綠色能源和可再生能源的應用可能是解決問題的途徑之一，但水資源問題仍然是不可忽視的挑戰。

小型語言模型（SLM）應該適用於現有的設備，並且其性能至少應該與解碼器相當。這意味著它們應該能夠在外部環境中良好運行。我們確實需要小型模型，但它們必須受到性能標準的限制。這些性能標準會受到我們所創建的軟體代碼的限制。小型神經網絡模型（SNM）的其中一個主要優勢是成本。由於這些模型較小，因此對資源的需求較低，



成本也相對較低。此外，它們的運行速度更快，效率更高。小型模型在定制化方面比較容易，因為你不必關心它們在多個任務上的表現，只需要關心它們在幾個特定任務。目前，小型語言模型的市場已經達到 80 億美元，預計未來將增長到 200 億美元。JNNR 市場預計將達到 3500 億美元，其中小型語言模型將佔其中的 5~6%。

儘管小型模型具有優勢，但訓練這些模型所需的能源是相當巨大的，這對環境造成了重大影響。最近在拉丁美洲的某些數據中心，由於水資源的短缺，甚至不得不拒絕繼續使用這些數據中心。關於能源問題，專家認為，綠色能源和可再生能源是解決問題的可行途徑。然而，水資源問題將繼續成為我們面對的一個主要挑戰。水是一種共享資源，也是運行數據中心所需的基本資源。我們需要乾淨的水來運行數據中心，這是一個非常迫切的問題。

**Table 1: Estimate of GPT-3's average operational water consumption footprint. "\*" denotes data centers under construction as of July 2023, and the PUE and WUE values for these data centers are based on Microsoft's projection.**

Location	PUE	WUE (L/kWh)	Electricity Water Intensity (L/kWh)	Water for Training (million L)			Water for Each Inference (mL)			# of Inferences for 500ml Water
				On-site Water	Off-site Water	Total Water	On-site Water	Off-site Water	Total Water	
U.S. Average	1.170	0.550	3.142	0.708	4.731	5.439	2.200	14.704	16.904	29.6
Wyoming	1.125	0.230	2.574	0.296	3.727	4.023	0.920	11.583	12.503	40.0
Iowa	1.160	0.190	3.104	0.245	4.634	4.879	0.760	14.403	15.163	33.0
Arizona	1.223	2.240	4.959	2.883	7.805	10.688	8.960	24.259	33.219	15.1
Washington	1.156	1.090	9.501	1.403	14.136	15.539	4.360	43.934	48.294	10.4
Virginia	1.144	0.170	2.385	0.219	3.511	3.730	0.680	10.913	11.593	43.1
Texas	1.307	1.820	1.287	2.342	2.165	4.507	7.280	6.729	14.009	35.7
Singapore	1.358	2.060	1.199	2.651	2.096	4.747	8.240	6.513	14.753	33.9
Ireland	1.197	0.030	1.476	0.039	2.274	2.313	0.120	7.069	7.189	69.6
Netherlands	1.158	0.080	3.445	0.103	5.134	5.237	0.320	15.956	16.276	30.7
Sweden	1.172	0.160	6.019	0.206	9.079	9.284	0.640	28.216	28.856	17.3
Mexico*	1.120	0.056	5.300	0.072	7.639	7.711	0.224	23.742	23.966	20.9
Georgia*	1.120	0.060	2.309	0.077	3.328	3.406	0.240	10.345	10.585	47.2
Taiwan*	1.200	1.000	2.177	1.287	3.362	4.649	4.000	10.448	14.448	34.6
Australia*	1.120	0.012	4.259	0.015	6.138	6.154	0.048	19.078	19.126	26.1
India*	1.430	0.000	3.445	0.000	6.340	6.340	0.000	19.704	19.704	25.4
Indonesia*	1.320	1.900	2.271	2.445	3.858	6.304	7.600	11.992	19.592	25.5
Denmark*	1.160	0.010	3.180	0.013	4.747	4.760	0.040	14.754	14.794	33.8
Finland*	1.120	0.010	4.542	0.013	6.548	6.561	0.040	20.350	20.390	24.5

Water consumption varies by location  
10-50 queries are answered per 500ml of water

### 圖七、水資源需求

當前常用兩種技術來構建小型語言模型（SLM），將現有的大型語言模型進行壓縮，提高效率。或從頭開始構建小型語言模型（SNM）：完全重新設計和訓練模型。小型語言模型（SLM）和大型語言模型（LLM）在某些情境下表現接近甚至超越傳統大型模型，特別是當使用高質量數據時。儘管小型語言模型不如大型模型強大，但在某些情況下，通過優化訓練數據和方法，仍能取得良好的效果。大型模型的結構仍然存在冗餘，未經最佳化，這為小型模型的提升提供了空間。例如，將世界知識壓縮至 7 億參數的模型可能就足以實現良好的表現，而無需 70 億參數，但數據的質量和使用方式比數量更重。

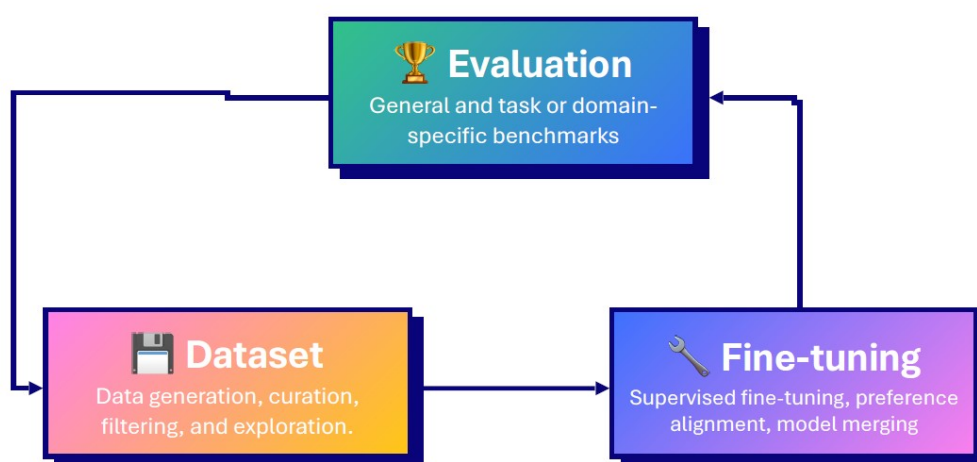
隨著技術的進步，模型的大小不再是唯一衡量標準，更多的創新方法能夠使得小型模型在性能和效率上超越傳統大型模型。不同的架構結合和更高效的算法將是未來改進的關鍵方向。這些技術的發展意味著，我們有更多的可能性來優化和應用小型語言模型，並且將來有望在硬體、演算法及計算效率等方面取得更多突破。

### 3、大型語言模型（LLMs）微調

LLM 會用大量的文本數據進行預訓練，主要目的是讓模型能夠預測問題的答案，這通常需要大量的數據。訓練模型能夠結構化地回答問題的階段。這一步需要專門的數據集，包括指令、問題和答案。進一步訓練，使用參考數據來不僅回答問題，還能以對人類有用的方式來回答問題。

微調並非總是最佳的解決方案，特別是在處理彈性問題時。微調需要一個適當的指令數據集，這樣模型可以根據問題的要求進行調整。評估框架對於確保解決方案的有效性

至關重要。這個框架不僅僅是為了衡量答案質量，還應考慮成本、延遲、吞吐量等因素，從而全面評估模型的表現。擁有一個良好的評估框架能夠幫助你確認解決方案是否達到了預期目標，並提供持續優化的反饋。數據集創建的三個關鍵維度：1.數據集中的每一個樣本是否正確？模型的回答是否真實且相關，尤其在處理數學或代碼等專業領域時，準確性是挑戰之一。2.數據集需要涵蓋多種不同的話題和使用場景，這樣才能幫助模型應對各種不同的情況，而不僅僅是特定領域的數據。3.數據集需要包含挑戰性較大的問題，而不僅僅是簡單問題。這能夠促使模型學習到更深層次的知識，從而提升其在複雜場景中的表現。



圖八、微調評估框架

訓練循環包括模型微調、數據集創建以及評估框架的反饋。在這個過程中，根據評估結果進行調整，並確保每一輪的迭代都能帶來改進。根據反饋，可能需要創建更多的數據或改善數據集，以便進一步優化模型。在構建和微調大型語言模型時，數據集的質量和多樣性、評估框架的設計、以及微調的選擇都至關重要。這些要素共同作用，幫助開發出更有效、更靈活的模型。重要的是，在開發過程中持續反饋並進行優化，以確保最終模型的表現符合預期。

數據格式在微調大型語言模型（LLMs）中的重要性，並區分了兩種類型的數據集，指令數據集（Instruction Dataset）包含了系統提示、來自用戶的查詢問題（指令）以及對應的預期答案。目的是訓練模型回答特定問題。偏好數據集與指令數據集的主要區別在於，這裡包含了選擇的答案和被拒絕的答案。目的是讓模型學會選擇正確答案，而不是產生不應該回答的選項。

另外通用型微調（如 Meta 的 Llama 3.1 聊天模型）是為了使模型能夠回答廣泛的問題，並且具備通用的能力。這需要大量的數據來涵蓋各種可能的問題和情境。而領域專屬微調其模型只需專注於特定領域，因此對數據的需求相對較少，成本也較低。任務專屬微調（如總結、翻譯）這些微調通常針對非常具體的任務，數據需求相對較小，可以專注於某一功能。這些任務屬於較狹窄的範疇，比通用型微調更具成本效益。

而數據可以是各種形式，如原始文本、文檔、提示或答案。通過生成指令和答案來轉換這些數據，這是數據處理的起點。而數據生成技術一般有反向翻譯（**Back-translation**），用來生成問題並讓模型學會回答問題、進化技術（**Evolution**）：當已有指令時，通過要求模型重新表述來增加問題的複雜性，以提升回答的深度。模型可以用來評估創建的樣本質量，並根據評分過濾掉不好的樣本。或清理可以基於啟發式方法或黑名單，也可以通過簡單條件進行過濾。過長的數據可能被移除，因為上下文窗口有其限制。

數據集需要轉換為精簡版本，尤其是在聊天模型的情境下，這樣模型能夠理解並按預期方式輸出答案。聊天模板（**Chat ML Templates**）訓練時會使用不同的模板，這些模板包含了系統提示（**System Prompt**）、用戶提示（**User Prompt**）和系統回答（**System Answer**）。這些模板幫助模型理解結構，並能夠按照指定的方式進行回答。

在進行微調時，數據的格式和結構至關重要。無論是指令數據集還是偏好數據集，合理的數據處理流程都能幫助優化模型的性能。針對不同的微調需求（如通用型、領域專屬或任務專屬），需要選擇合適的數據集和處理方法。有效的數據過濾、清理和重複數據去除可以提高訓練效率，並使最終的模型能夠按照預期方式生成有用的回答。

技術和方法能夠幫助提升微調過程的效率和質量，無論是針對通用型模型、領域專屬模型，還是任務專屬模型。模型合併和自定義基準測試是實現高性能微調的重要工具。這些信息能幫助你更好地理解如何進行有效的監督式微調，並選擇最適合你需求的技術和方法。

#### 4、優化封閉網路 LLM

AI 性能的提升有相當高的期望，並且已經投入了大量資源進行優化。GDG4 和 Turo 的技術聽起來很有前瞻性，尤其是在記憶體利用和量化技術方面。能夠達到與 GPT 類似的效能，並且在壓縮型環境中提供超過 280GB GPU 效能，這是一個非常重要的突破。

雖然投入已經帶來顯著的效能提升，但持續關注開放模型和高效運算的結合，可能進一步加速進步。開放源碼技術在資料處理、模型整合等方面的應用，可以將成本降到最低並且更具靈活性。你們的資金應該在提升基礎設施的同時，關注模型創新和可持續優化。

將技術整合到私有雲中的計劃是非常有遠見的。透過開源和訂閱模式的混合方式來提供定制化服務，這對於企業和大規模應用來說是非常合適的策略。基於現有框架，將它們作為 API 進行容器化是一個不錯的選擇。這樣可以為多種應用場景提供支持，且便於跨平台的使用。應該進一步探討如何在其他領域（如推薦系統、搜索引擎）中應用這類技術。

將參數從 405 億擴展至 3,371 億的目標是值得挑戰的，這將對計算資源和數據處理提出更高要求。你可以考慮使用分布式訓練（比如 Horovod）來分擔訓練過程中的計算負擔，並利用更多的計算節點加速模型的訓練。建立標準化的自動化基準測試，這將有助於比較模型性能。你可以選擇多個領域的標準數據集，並針對不同任務設計特定的基準測試。例如，針對韓文處理的基準測試可以與多語言文本生成的基準測試結合使用，從而全面評估模型的多樣性和性能。

除了傳統的基準測試外，RNA 風格的用戶偏好評估也是非常重要的，尤其是在聊天機器人等應用中。通過收集用戶的偏好數據，並將其與其他模型進行比較，你可以獲得更具實際意義的性能評估。

模型合併技術（例如線性插值），這是提升模型效能的有效方法。將多個模型的強項結合，能夠達到更強的綜合性能。在你的情境中，這類技術應該對韓文領域的特化模型有所幫助，尤其是在多領域模型整合的背景下。

在公共雲中，服務結構表（例如 HWP 表）存在大量數據，這些表格資料與訂閱模式密切相關，並且這些需求在全球範圍內呈上升趨勢。

如何有效管理數據的攝取、存儲、處理和監控成為關鍵問題，尤其是在面對大量數據和不同模型的情況下。開發開源工具來簡化 LLM 的處理過程，這樣可以降低開發門檻並提高處理效率。例如，使用 Apache Kafka 或 Apache Flink 進行高效的數據流處理，來解決大量數據攝取與即時處理問題。

如提到的 Mosaic 資料庫模組，這種模組化的數據管理系統能夠有效地存儲並管理數據中的參數、變數和相關資訊，並記錄開發歷史，這樣可以簡化開發和調試過程，提高服務的運行效率。當同步次數增加時，對 GPU 的需求也會大幅上升。這會造成資源瓶頸，尤其是處理大型語言模型時。前端設置負載均衡系統，可以將流量分散到多個伺服器，並在需求量增加時動態調用更多的 GPU。引入類似 Misture 的微服務，可以在工作負載增加時自動平衡負載，這樣能夠優化硬體資源的利用。利用 GPU 的升級來提高計算能力，並根據實時需求動態調整 GPU 的數量和配置，從而達到高效處理模型的目標。

引入 MLOps 平台，如 Kubeflow 或 MLflow，來實現多模型的管理與測試。這些工具可以幫助自動化模型選擇、訓練與測試過程，並自動調整參數（如批量大小、學習率等），從而提高訓練效率。針對大規模模型的訓練，可以採用 分布式訓練技術（如 Horovod），通過多台機器或多 GPU 來加速模型訓練。通過 自動擴展（Auto-scaling）技術，根據實時流量或需求的變化動態調整雲資源，無論是 GPU 還是其他計算資源。這樣可以有效處理流量波動，保證服務在高峰期也能穩定運行。利用 Prometheus、Grafana 等工具來監控伺服器的運行狀態，及時發現問題並作出調整。

將數據處理轉移到 邊緣計算 節點，尤其是對於需要實時反應的應用，邊緣計算能夠有效減少數據傳輸過程中的延遲，從而提高服務的反應速度。使用 推斷機制（如 ANT）來動態推測外部數據的需求並自動調整系統設置。這樣，儘管外部數據的訪問可能有限，但內部系統仍能根據外部條件進行動態調整，降低數據外洩的風險。

**反應代理（Reaction Agent）**：根據 System 1 和 System 2 的思維方式，當遇到簡單問題時可以直接處理，而複雜問題則需要深入思考。這樣的代理系統可以根據問題的難度自動選擇處理策略，並在推理過程中動態調整。這樣能夠提升系統在複雜決策過程中的靈活性。

通過 AgentTip 技術，可以根據不同的問題類型選擇最適合的工具進行處理，這樣可以更高效地解決各種業務需求。例如，當面對具體的業務場景時，系統可以通過自動選擇 查詢工具 或 摘要工具 等來提高答案的準確性和運行效率。

利用 詞彙記憶（Word Memory），來提高系統在多任務環境中的表現。這可以幫助系統在面對大量不同問題時，能夠有效地記憶和運用關聯信息，從而提高整體的運算精度和效率。

將不同問題的處理需求進行合併，利用共享的知識來解決多個任務，這樣的技術能夠幫助提高系統在面對複雜場景時的學習能力。

公有雲環境中的挑戰涉及多個方面，包括數據處理、硬體資源調整、模型測試和性能優化等。利用負載均衡、自動擴展、微服務架構、推斷機制以及多任務學習等技術，我們能夠在不斷變化的業務需求和技術挑戰中，實現高效、穩定的服務運行。

## 九、 AI 代理來協助產品開發。

AI 的運作方式發生了根本性的改變。以往開發 AI 產品需要深厚的技術背景和編程知識，但現在只要你有一個清晰的想法、正確的框架和結構化的方式，即便沒有技術背景，任何人也能開發 AI 產品。隨著生成式 AI 的發展領域變得非常廣泛，數以千計的公司已經投入其中。其應用範圍包括，如文本生成、程式碼生成、語音生成、圖像生成、影片生成、3D 物件生成及數據增強（特別是用於訓練 AI 模型的合成數據）等

可能會看到一些由單一個人運營的億萬美元公司出現。聽起來可能很瘋狂，但這正是 AI 時代的趨勢，因為每個人的技能都在不斷提升。重點是如何利用 AI 技術來提升研究能力，如何高效地利用聊天機器人以及如何程式開發環境中應用這些工具。

過去，前端開發是最可怕的領域之一，因為這需要專業的前端開發人員。可是，隨著技術的發展，現在即便沒有專業技能，我也能輕鬆進行這些工作。類似的情況也發生在資料庫開發和雲端開發領域。舉個例子，假設我想開發一款 iPhone 應用程式，我會問：「開發一款 iPhone 應用程式需要哪些關鍵組件？從規劃設計到開發再到部署到 App Store，過程中需要注意哪些問題？」，這其實是個非常龐大的問題，但如今我們能夠通過 AI 工具來幫助簡化這一過程。以 Complexity AR 為例，它能夠幫助我快速檢索網絡上的相關資料，並以結構化的方式向我展示應該考慮的關鍵問題，這些資源可以讓我輕鬆了解開發過程中的每一個細節。

有些工具可以幫助我們快速理解複雜的領域，並做出決策。最近，O1 這個工具也取得了巨大進展。它不僅能夠為我們提供高水準的算法解決方案，還能夠協助開發者生成更加精準的代碼。更強大的是 Cursor AI，它是我目前最喜愛的工具之一。它能夠根據簡單的提示，幫助我們快速改進腳本，甚至無需編寫代碼，只需給出指令，AI 便會完成所有步驟。以往開發雲端應用程式需要專業的基礎設施工程師來處理各種計算資源、儲存問題等。但現在，即便是這樣的工作也可以通過 Copilot 和其他 AI 工具來完成。只需要簡單的指令，AI 就能夠自動檢查並執行任務，例如啟動虛擬機器、處理存儲資源等。

Martin Luzio 和我的團隊開發了一個框架，旨在幫助每個人都能夠輕鬆編寫代碼。這個框架的核心是 AID Framework，包括以下步驟：Ask（詢問）：定義你最小可行產品（MVP），向 AI 提出問題以幫助其理解你的需求、Identify（識別）：透過 AI 的多輪對話來識別所需的技術選型，並開始進行資料搜索和方案選擇、Document（文檔）：當我們對產品的需求有了共識後，可以要求 AI 自動整理成一份文檔，作為開發的依據。及 Execution（執行）：開始實際開發工作，AI 會協助你完成編碼、測試和部署，這個框架能幫助我們從一個簡單的想法出發，逐步實現我們的 AI 產品。

「我不懂技術，能不能做出 AI 產品？」那麼現在，答案是：完全可以！未來，AI 技術將讓我們的每個人都成為開發者，幾乎所有的工作都能透過簡單的對話來完成。今天



的分享只是冰山一角，我相信 2025 年我們將看到更加驚人的變革，這將會是個充滿創新和機遇的時代。

## 伍、心得與建議

### 一、心得

AI 產業的快速發展，未來將進一步改變各行各業的運作模式，尤其是在金融和風險管理領域。於傳統製造業隨著物聯網等新技術及 AI 的應用，將會深入到更多的生產線和業務流程的變革。為提升運營效率、客戶服務以及解決技術的挑戰，將會成為企業導入 AI 的原因。企業在 AI 投資和發展中面臨的挑戰，尤其是高成本和技術運營上的難題。企業應該依賴現有的解決方案，設立標準，並進行合作以降低成本並提高效率。此外，儘管 AI 技術的普及需要時間，但隨著市場對其需求的增長，企業應該逐步適應此技術的發展。

對於管理者和組織領導者來說，導入 AI 和自動化流程，整體是件好事。但部署 AI 應用時，哪些工作可被 AI 取代？哪些工作可藉由 AI 工具提升相關效率？哪些工作是因應 AI 的導入而創造/新增出來的？在規劃導入前，或須先行盤整定義，若涉及組織業務流程的更改，更應預為規劃因應。

基本上，我們的目標應是創建能夠支持自我提升的 AI 系統，幫助使用者增強自身能力，如增強創造力、促進社交，而非削弱人類的核心價值、取代使用者。因此預為規劃提供「在職學習」途徑，協助員工轉型，是不可或缺的成功關鍵要素。

AI 科技的「適用領域」是必須注意的事情，以本會資訊處要導入應用 AI 言，建置 LLM（大型語言模型）無疑是不切實際，Meta 講者在「小型語言模型與神經網絡」中提及：小型語言模型相比大型模型，在隱私敏感領域（如國防、法律、醫療等）有明顯優勢，可在本地部署，避免將數據傳送到外部伺服器，從而確保數據安全。且小型模型的運行成本較低，對硬體要求也較小，使其更具可操作性。雖然在專業領域內對小型神經網絡模型（SNM）的定義尚未有共識。重點在於模型之間的區別不僅僅在於它們的大小，而是取決於它們是如何訓練的，或者它們的使用目的。小型模型在定制化方面比較容易，能夠更加專注，在幾個特定任務上表現能更良好。反之，企業應該專注於分析現有的解決方案，並確定哪些解決方案最適合他們的需求。這樣可以避免在自建模型上浪費過多資源。因此，本會資訊處要導入應用 AI，或可站在巨人的肩膀上看世界，直接使用預訓練好的大型語言模型，再落地界接就地端的 RGA 模型，將更符合成本效益。

研討會中提到搜尋、視覺創意及人形機器人導入 AI 科技，站在道德高點的角度，很多人主張不能單純地將 AI 的失敗視為「學習過程」或「不可避免的錯誤」，因為一旦發生錯誤，影響和傷害非常嚴重。相對於那些高風險的領域，「創造性」和「表達性」的任務，如藝術創作或行銷文案撰寫，對 AI 的失敗就很寬容。在這些領域中，即使 AI 創作出不受歡迎或者品質低劣的作品，後果就只是別人不滿意，或是市場接受度很低，不會影響到人們的安全或健康。另外 AI 是一個基於數據和演算法的工具，當 AI 作出決策時，它依賴的是「過去」的數據和經驗，這種依賴可能會導致判斷上的「偏誤」，也就是 AI 的決策只是針對「過去」和「現在」現象的反映，而缺乏對「未來」或更崇高理想的追求。

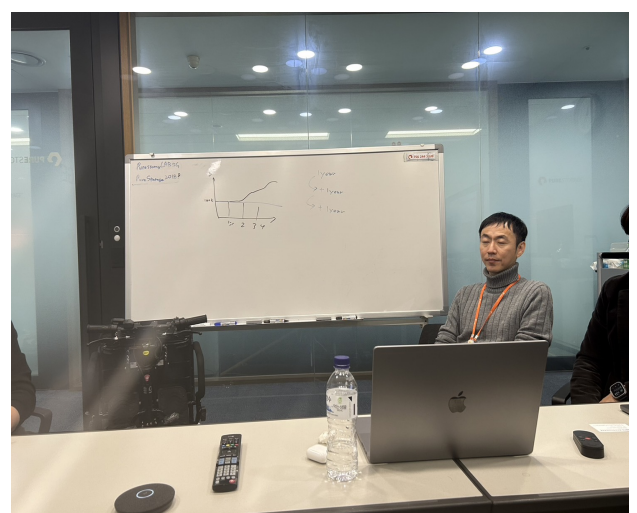
對於個人工作來說，理解現在 AI 的發展趨勢，以及未來可能會面對的一些轉變。例如類似「我會被 AI 取代嗎？」這種問題，將會發現「人的價值」開始轉變，我們思考的是如何轉型和適應，甚至在必要的時候主動介入、操作方向盤。因在 AI 持續蓬勃發展的未來，會被 AI 取代的，是無法轉型和適應新價值的人。

## 二、建議

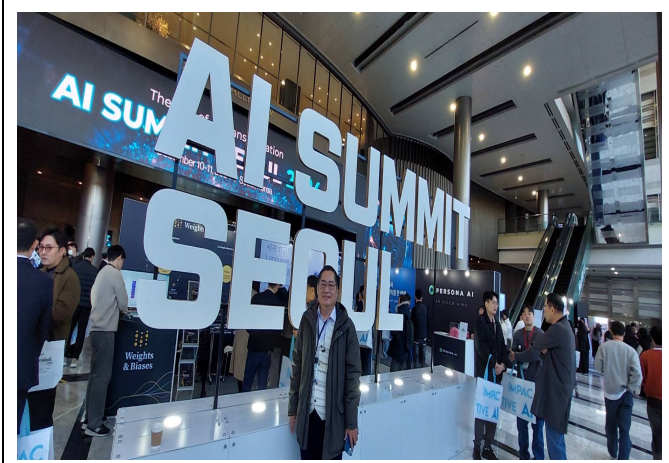
1. AI 技術的發展，一些模型不僅處理文本，還能處理圖像、語音、視頻等多媒體數據型式，在視覺創作和故事創意表達，具備多模型態，可提供於多媒體製作協助機關業務或形象行銷。
2. 面對像 LLM 這樣的高成本技術時，依賴於已有的解決方案可能會是一個更具成本效益的選擇，在不同的應用市面上也有一些商業化成熟產品如微軟 copilot 及一些翻譯或與眼處理等通用性產品，建議機關直接導入這些產頻避免重複開發。
3. 研討會中演講者的實戰經驗分享，結合本(113)年度幾次與國家高速網路公司 TAIDE 模型訓練團隊幾個議題的討論，促使我們有下列想法：未來本會要落地導入具有公文生成、立委模擬問答生成、新聞稿生成、人民陳情案回覆生成功能的「TAIDE 模型」，為了有較佳的執行績效與彈性，似乎朝分別建置各不同功能落地端 RAG 的方向進行，是個可考量的方法。
4. 有沒有需要將現有大型語言模型進行壓縮，轉構建為小型語言模型以提高效率後再落地，以本會現行之資源架構，似乎可朝先直接將落地端 RAG 界接後端 TAIDE 模型，於測試環境上測試作業效能後，再予決定。值得提醒未來團隊在導入落地 AI 模型時：數據的質量和使用方式比數量更重要。如何藉由不同的架構組/結合和更高效的算法，將是未來改進的關鍵方向。
5. 過去公司可能只依賴一兩個大模型，但如今使用多個模型變得越來越普遍。或許，當本會有落地導入就地端「TAIDE 模型」的經驗後，可朝著思考是否有機會將會內的外顯知識(或為既有之知識庫)資料，將其轉換成向量空間，通過自然語言嵌入 API 來塑模成可供檢索的 AI 服務，即藉由提示引擎工具，且將相關問題轉換為向量後，語言模型就能根據這些向量找到相關資料來回答問題，提供同仁參考。



陸、附錄-研討會議程及資料



Purestorage 參訪







## 研討會

Day 1   Dec 10		
Time	Agenda	Speakers
08:00 - 08:50	Kyocera Panel Market Trend: AI Investments, Ecosystem, and Technology From the Perspective of a Silicon Valley Investor Kyocera Panel Market Trend: A Look at the Present and Future of AI Agents Kyocera Panel Opening Talk: AI Next 10 Years: Beyond Hype, into Inevitable Possibilities	Jay Eum CEO of T Ventures, Silicon Valley Investor, (Former) McKinsey Kim Chan-yeop Professor / Korea University, former Vice President of Samsung Electronics Choi Joo-yeon, Professor / KAIST
09:50 - 10:15	Headline Trend Human-Centered AI in 2024	Ben Shriklesman, Human-Centered AI Author, Professor, University of Maryland
10:15 - 10:45	Morning Break	
10:45 - 11:20	Kyocera Special AI Product Strategy AI Product to Scaling up	Ben Barone-Nugent, Principal, Content Designer, AI Group Lead   Carista Jonathan Elliott, Content Designer, Facebook Design Systems   Meta
11:30 - 12:00	Kyocera Talk: Enterprise AI Strategy for AI First Enterprise	Song Yong-jin, Vice President, Global Business Jung Hee-jin, Senior Vice President, Hardware Systems, former CEO & COO of Samsung Joo Seon-yeop, Executive Vice President, Korea Retail Division Jung Dong-ik, Professor, KAIST, former CEO & COO of Samsung Minsook, Senior Advisor, former CEO of Samsung
12:00 - 13:20	Lunch Time	
13:30 - 13:55	AI in Education How GenAI is Revolutionizing Creative Expression Learner's Perspective APAC Creative Tech Leader Publicis Group	AI Transformation Why do AI projects fail? Conditions that reduce failure Jung Dong-ik, Professor   KAIST Kim Jaeyoung, Graduate School of AI
13:55 - 14:30	Gaming Industry Applications of AI Technology in the Gaming Industry Kim Min-jae, Center Director   NCsoft	AI Innovation Taking Agentic AI systems to production Adeo Solutions Country Manager   Weights & Biases
14:30 - 15:05	Medical Industry Building the largest biomedical LLM in the World Mettis Presepal, Professor   University of Florida Yi Qun, Professor   University of Florida	Productive Analytics Predictive AI is Revolutionizing Productivity: How Small Businesses Can Apply It Jung Doo-hee, CEO   Inspirev AI
15:05 - 15:40	15:05 - 15:20 Ad Industry AI Breaking Boundaries in Video Production: New Possibilities Unlocked by AI Lee Chang-woon, Director of Research   Leader Korea	DevOps DevOps at Netflix Talesi Chopra, Senior Software Engineer   Netflix
15:40 - 16:05	15:20 - 15:40 Healthcare AI Innovative Applications in Medical Diagnosis Sung Nak-woon, VP   PERSONA AI	AI Agent Agent AI: Navigating the Next Era of Autonomous Systems Kazuhiko Hayashi, Field Marketing Director, APJ   Sambanova
16:05 - 16:35	Virtual LLM On-the-ground GPT Innovation: Full-Scale Adoption Story in a Consumer Goods Company Choi Young-won, Director   Donggwan Industries	AI Transformation Money-Making AI: How Korean Leverages AI to Analyze the Beauty Market Kang Yoon-hye, Executive Director   Korea Korea Choi Young-min, CEO   Unisound Lab
16:35 - 17:10	Manufacturing Industry AI Applications and Real-World Use in the Manufacturing Industry Jung Seung-yeon, Managing Director   Doosan Energy	AI Product Beyond B2B: Developing AI Products That Deliver Real Value! Story Ben Barone-Nugent, AI Group Lead   Carista, former Google Gemini
17:30 - 17:40	LLM & Knowledge Graph AI Transformation in a Digital Platform Company: The Journey and Lessons with AI Agents Joo Hyung-min, General Manager   Wersid Lab	AI Start-up Showcase Generative AI Market Solution Case: Lee Jung-yeon, CEO   NewTune Generative AI Safety, Diagnostics and Vulnerability: Defense Solution Case: Yoo Sang-yeon, CEO   Iron Intelligence Conversational AI Agent Solution Case: Specialist for Commerce: Park Ahyun, CEO   Wisadea

Day 2   Dec 11		
Time	Agenda	Speakers
08:00 - 08:25	Opening Talk Opening Talk	Sam Han, Head of AI & Data   The Washington Post
08:25 - 10:00	Kyocera Panel Trend AI Adoption Journey: Conditions for Successful Enterprise AI Adoption	Nikhil Dwaraknath, Group Head of Data   Grab Mehmet Firatoglu, AI Scientist   Mercedes-Benz AG Sergio Ochoa, MD of AI/ML and Quantum   Moody's Analytics Mehmet Firatoglu, MD of AI/ML and Quantum   Moody's Analytics Gibran & Co. University Graduate School of Business Administration
10:00 - 10:20	Headline: Gen AI Building and Scaling a New Generative Media and Entertainment Company	Cristoforo Valentini, Co-founder and CEO   Runway
10:20 - 10:50	Morning Break	
10:50 - 11:30	Headline: Robotics The Current State of Humanoid Robots	Peggy Johnson, CEO   Agility Robotics Conversation with Samsung I&D Senior Reporter - Asia Artificial Intelligence, Bloomberg
11:30 - 12:10	Headline: AI Chips AI Supercomputing and Services Centered Around HBM	Kim Jaung-ho, Professor   KAIST
12:30 - 13:30	Lunch Time	
13:30 - 14:05	Media Industry AI Adoption Journey at The Washington Post Sam Han, Head of Data & AI   The Washington Post	LLM Trend Understanding the Evolution of LLM-Based AI Technologies: RAG, AI Agents, and Future Outlook for 2025 Choi Yoon-seok, Tech PM   Microsoft
14:05 - 14:40	Financial Industry From Proof of Concept to Production: Battle-Tested Insights from 100+ Enterprise GenAI Deployments Nikhil Dwaraknath, Group Head of Data   Grab	LLM Model LLM: Small Models Modern Techniques Soumya Batra, Tech Lead, Applied Research Scientist, former Meta
14:40 - 15:15	Financial Industry Evaluating LLM Performance in the Financial Industry Choi Young-jun, Team Leader   KakaoBank	LLM Tech LLM Fine-Tuning and Evaluation Maxime Labonne, Head of Post-Training   Liquid AI
15:15 - 15:50	AI Translation How to Compete with Google, Microsoft and OpenAI Christopher Osborne, VP of Product   DeepL	LLM Implementation Strategy Building Private LLM-based Agents: RAG in Closed Corporate Networks Joo Chul-hwi, Director   AFISCA
15:50 - 16:20	Delivery Industry GPT-powered NLP Innovation Story: What Menu Do Customers Really Want? Oh Hyun-joon, Team Leader   Wooja Brothers	AI Infra AI & Gaming Infrastructure of the Future Mark Ryan, Co-founder & CEO   Anthr
16:20 - 16:55	HR Industry & AI A CEO's Guide to AI: How I Can Make Your Organization More Efficient Lee Jung-ho, Manager   Korea Back-Pose Consultant Lee Seung-ho, Professor   Korea University Kim Jin-ik, Executive Director   GS	LLM Optimizing Large Language Models (LLMs) for Scalable and Ethical Enterprise Applications Soumya Batra, Tech Lead, Applied Research Scientist, Maxime Labonne, Head of Post-Training   Liquid AI Martin Meisel, Gen AI Leader   GenerativeAI.net
16:55 - 17:30	Edu Tech How is AI transforming EduTech and the learning ecosystem? Kim Deok-jong, Professor   Soekyeong Women's University	AI & Regulation Current Status and Future Outlook of the EU AI Act: Strategic Preparation from a Business Perspective Lee Dong-jun, Lawyer   Hogan Lovells International

## 研討會行程