數位再轉型：用客製化大模型解決科技問題

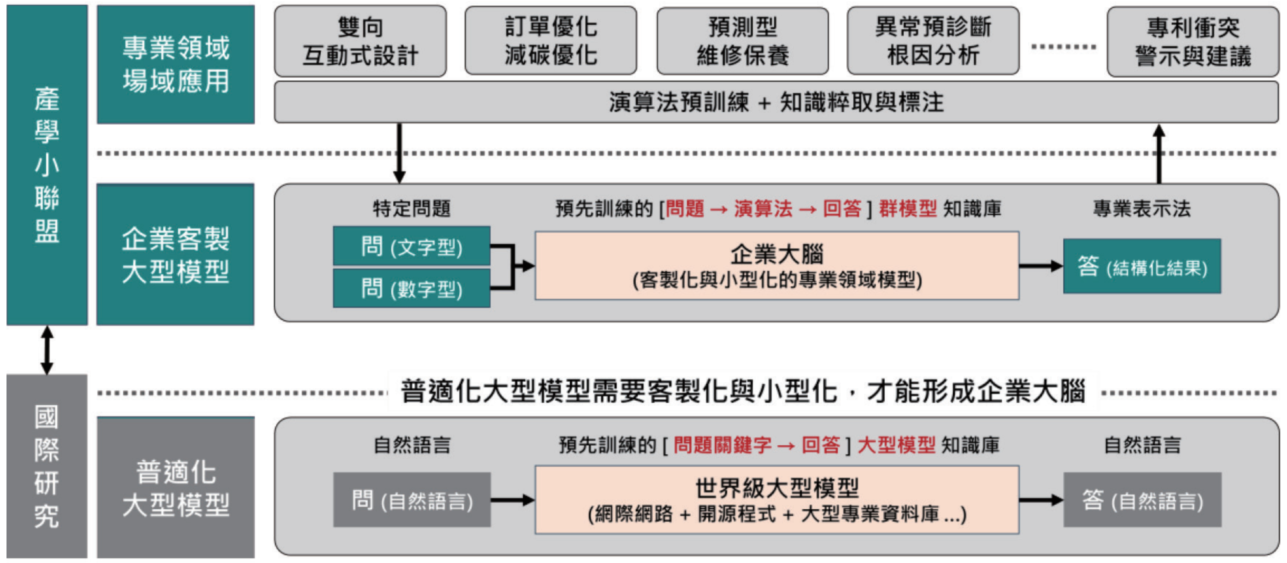
文／杜懿洵

自 2022 年底 OpenAI 發布 GPT-3.5 後，因為能夠對話與更能夠接近人類思考方式等特性，正式引爆大型語言模型 AI 熱潮；然而，GPT-3.5 之所以能引發熱議，在於從第一代開始不斷砸下的重金與資源，除了演算資源之外，GPT-3 模型參數已高達 1750 億，而這也是自 GPT3.5 開始，能將訓練從單向的資料提供，轉向對話模式的關鍵。2023 年 3 月，OpenAI 緊接著發布 GPT-4，更支援視覺輸入、圖像辨識，在強大的參數訓練之後，AI 的重點也快速的轉向提升利用現有數據的能力上，而各種企業應用研發也如雨後春筍般興起。

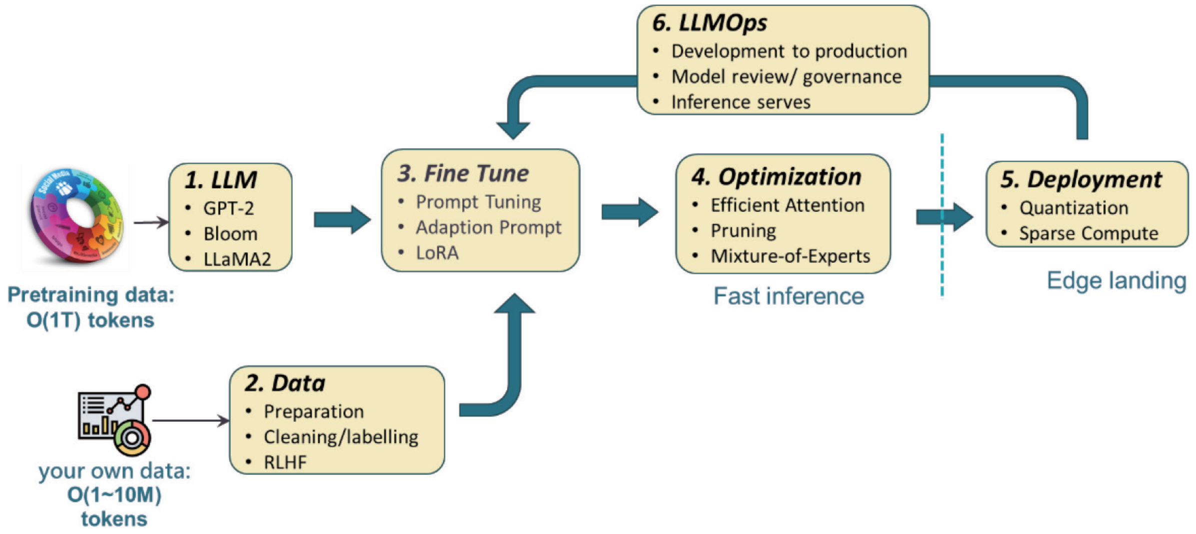
隨著 GPT 的應用風起雲湧，AI 技術中的大型語言模型（LLM）不僅提供了強大的自然語言處理能力，更潛在地改變了各個產業，從智慧製造到永續發展，從知識管理到企業大腦的構建。為了分享國內產、學、研專家在大型模型的研究成果以及目前標竿企業實際的落地應用，國立陽明交通大學人工智慧系統檢測中心與國家高速網路與計算中心、台灣人工智慧協會共同合作，於 2003 年 8 月 18 日聯合主辦「2023 LLM 產學技術交流會」，以「AI 大型模型趨勢發展探討」、「永續經營與智慧製造」、「AI 落地應用分享」三大主題，邀請近 30 位產、學、研的領先專家分享見解和經驗，現場並有 27 項技術實體的展示攤位，期待透過產、學、研的分享與交流，一起攜手為開發出台灣自己的 GPT 而努力。

然而，發展台灣大型語言模型雖已成共識，但訓練一個客製化大型生成式 AI 模型，不但需要大量時間與金錢成本，更需要來自產學研的三大核心能力：「學術界的演算法發展能力」、「企業的知識庫整合能力」，以及「專業研究機構的高效計算能」，才能與國際巨頭競爭。有鑑於此，陽明交大倡議成立「大型模型產學小聯盟」，共同建立一個可分享又具有私有化保密機制的大型模型。

「大型模型產學小聯盟」的成立，除了在打造更符合台灣人使用習慣的 AI 平台之外，更期待能讓大型語言模型從普適性的通用模組進入特定領域的產業，進行客製化的應用，以及讓企業能以大型模型的縮小版，做出自己的客製化知識管理系統，訓練出自己的「企業大腦」，為未來開發出更多產業新機會（圖一）。



資料來源：陽明交通大學 陳添福教授



圖二 有效率的大型語言模型開發流程

然而，若要達成「大型模型產學小聯盟」的目標，亟待企業界在資深專家結構化與非結構化知識萃取與標注、私有資料、大型模型合成數據 (synthetic data) 等知識庫的整合，讓模型能在企業落地實踐中學習。

私有資料的清理與整合是大型模型成功的基礎

陳添福教授表示，要做出一個有效率的模型運用，一定要做好私有資料的清理與整合 (步驟 2)，接著才是模型的預訓練 (Pre-trained)、微調 (Fine Tune) 以及優化 (Optimization)，但目前 LLM 存在著「缺乏企業私有化的資料」，尤其是技術資料，以及預訓練資料過時的缺點，因此，期待能與企業合作，進行私有資料的清理與整合，讓 LLM 能得到正確、專業的互動結果。

應用 Retrieval-based 的大型語言模型框架整合私有資料

而在大型模型合成數據部分，目前主要是以 Retrieval-based 的大型語言模型框架，進行企業私有知識庫檢索與生成式模型整合。Retrieval-based 框架的檢索擴增生成理論 (Retrieval-Augmented Generation, RAG)，不但能快速且精準地從資料庫中檢索相關的參考文獻、提供簡潔和易於理解的答案，也考量資料安全和權限管理的問題，針對不同權限層級的資料進行隔離和分類，讓使企業能夠保護機密信息。交流會中，陽明交大也展示 Retrieval-based 框架應用的 RISC-V 知識小助手作為實際參考。

老師傅經驗的萃取與標注與 AI 客製化插件

至於如何將資深專家 / 老師傅的經驗與智慧，進行結構化與非結構化知識的萃取與標注，向來是所有 AI 關注的重點，而此次交流會中，中研院孔祥重院士也分享了從影像語言模型萃取老師傅經驗的案例。而針對客製化經驗萃取開發專業插件，也是針對已經有一定發展程度的 AI，一個值得關注的議題與相對務實的做法。交流會中，優智能吳浩平技術長以及多位專家也分享了面對企業生產情境多樣，以客製化的經驗萃取插件，將企業員工的個人知識轉化為組織可以再利用的 AI 工具的選擇與應用，提供大家進一步思考。

「2023 LLM 產學技術交流會」不僅有產學研的專家給予技術知識，現場也聚集了許多尋求解決方法的產業人士，多元化的參與組成，豐富了活動的內容並促進跨領域的交流。總體而言，除了期許 LLM 的發展上，能打造更符合台灣人使用習慣的平台，更希望是在企業可以負擔的情況下，打造一個兼顧資安與智慧的知識庫，為未來的 LLM 發展和應用打下了堅實的基礎。如欲詢更詳細資訊，請洽 myLLM.tw 網站。

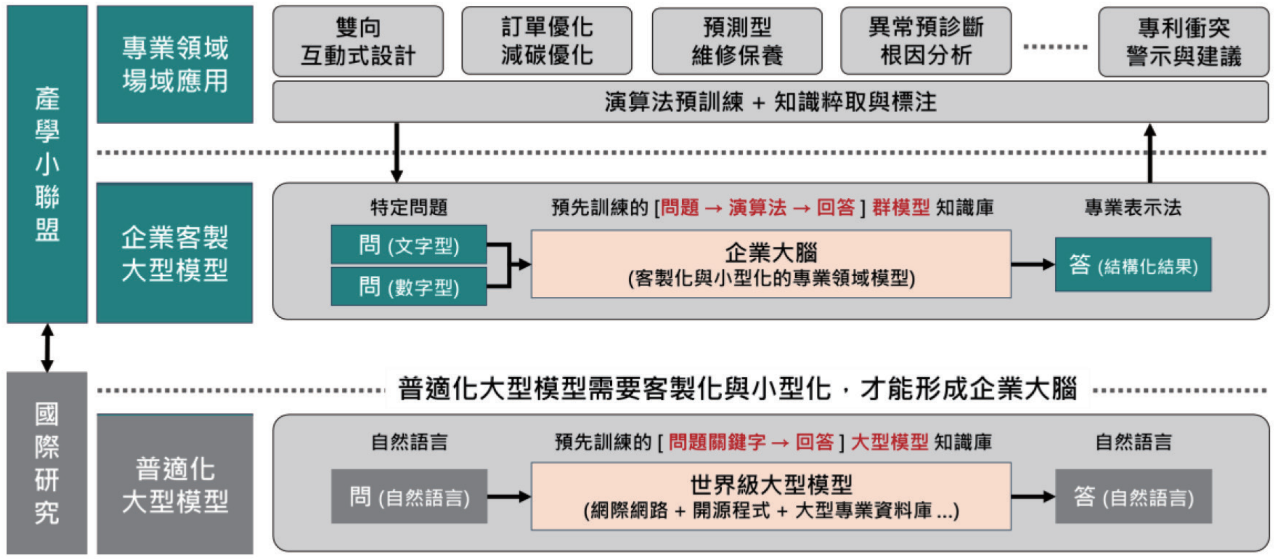
Digital Transformation: New Approach for Problem-Solving with Custom Large Models

OpenAI's release of GPT-3.5 at the end of 2022 has sparked a surge in the popularity of large-scale language models for generative AI, thanks to its ability to engage in conversations and closely approximate human thought processes. The fervor around GPT-3.5 arose from the significant investment of resources poured into its development, since the first generation. GPT-3 boasts 175 billion parameters, which played a key role in shifting the training focus from unidirectional data provision to dialogue mode. In March 2023, OpenAI quickly followed up with the release of GPT-4, which supported visual input and image recognition. After intensive parameter training, AI shifted its focus towards enhancing its ability to utilize existing data, resulting in the proliferation of diverse enterprise applications and research developments.

The use of Large Language Models (LLMs) in generative AI is rapidly expanding. These models possess powerful natural language processing capabilities and have the potential to revolutionize various industries. They are utilized in smart manufacturing, sustainable development, knowledge management, and corporate intelligence. To disseminate the findings of domestic industry, academia, and research experts in large-scale models, the AI System Benchmarking and Tuning Lab at National Yang Ming Chiao Tung University, in collaboration with the National High-Speed Network and Computing Center and the Taiwan Artificial Intelligence Association, jointly hosted the "2023 LLM Technology Exchange Conference" on August 18, 2023. The conference had three main themes: "Exploring the Latest Trends in AI Large Models Development," "Sustainable Operations and Smart Manufacturing," and "Sharing the Deployment of Artificial Intelligence in Real-World Practice." About 30 renowned experts from industry, academia, and research were invited to share their insights and experiences, along with 27 show booths on-site. The goal was to collaboratively share and exchange knowledge among industry, academia, and research communities to coordinate efforts toward developing Taiwan's indigenous GPT.

There is a consensus in Taiwan about the development of large language models. However, creating a customized large-scale AI model requires considerable time and financial investment, as well as three core capabilities from academia, industry, and research: “the academic community must have algorithm development skills”, “the enterprises must have the capacity to integrate knowledge”, and “the professional research institutions must have efficient computing capabilities”, which are crucial for competing with international leaders in this field. Therefore, National Yang Ming Chiao Tung University proposes the formation of the "Large Model Industry-Academia Small Alliance" to collaboratively build a large model that features both shareable mechanisms and privacy protection.

The establishment of the 'Large Model Industry-Academia Small Alliance' not only aims to develop AI platforms catering to the preferences of Taiwanese users but also intends to facilitate the evolution of large language models from universal, general-purpose modules to industry-specific applications. This evolution enables enterprises to develop their customized knowledge management systems using scaled-down versions of large models. By training their own 'enterprise brains,' this initiative sets the stage for the emergence of numerous industrial prospects in the future, as depicted in Figure 1.



Source: Professor Tien-Fu Chen, National Yang Ming Chiao Tung University.

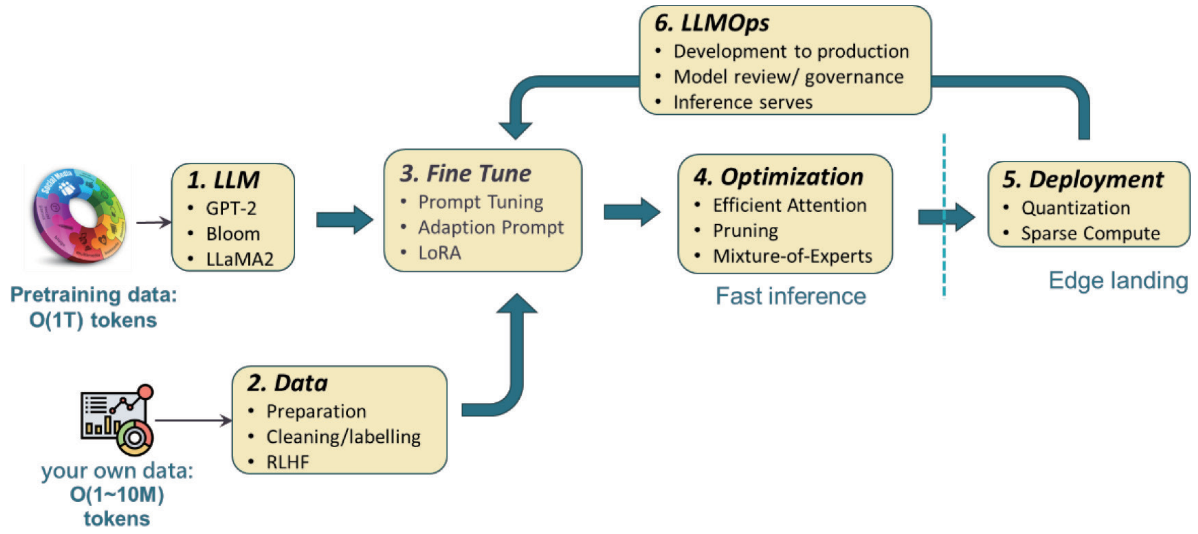


Figure 2 Efficient Development Process of Large Language Models.

To achieve the goals of the 'Large Model Industry-Academia Small Alliance,' it is crucial that businesses promptly integrate annotated information derived from both structured and unstructured data by experienced professionals, private data, and synthetic data for large-scale models into knowledge bases to facilitate model learning during practical deployment in enterprises.

Private data cleaning and integration is the foundation of the success of large-scale models

Professor Tien-Fu Chen has highlighted the significance of data cleaning and integration as a crucial step toward creating an efficient model. This process comes as Step 2, followed by pre-training, fine-tuning, and model optimization. However, the current Large Language Models (LLMs) suffer from the lack of privatized enterprise data, specifically technical data, and the issue of outmoded pre-training data. Therefore, there is growing anticipation for collaborative efforts with enterprises toward private data cleaning and integration, ensuring that LLMs can deliver precise and professional interactive outcomes.

Figure 2 Efficient Development Process of Large Language Models.

Integrating private data using Retrieval-based large language model frameworks

In terms of data synthesis with large-scale models, the retrieval-based framework of large language models is primarily utilized to integrate enterprise private knowledge retrieval with generative models. The Retrieval-Augmented Generation (RAG) framework within the retrieval-based framework enables quick and accurate retrieval of relevant references from databases to provide concise and understandable responses. Furthermore, it addresses concerns regarding data security and access control management by dividing and categorizing data based on different authorization levels, empowering enterprises to safeguard confidential information. At the conference, a practical application of the retrieval-based framework was presented by NYCU as a RISC-V knowledge assistant for reference.

Tacit knowledge extraction, annotation, and AI-customized plugins

The focus of all AI attention has traditionally been on harnessing the experience and wisdom of senior experts or mentors to retrieve and annotate both structured and unstructured knowledge. At the meeting, Academician Dr. Hsiang-Tsung Kung from Academia Sinica presented cases illustrating such information extraction using image-language models. The development of plugins based on customized experience is a valuable and practical strategy for AI systems that have reached a certain level of advancement. During the meeting, Dr. Haopin Wu, Chief Technology Officer of Goedge.ai Inc., and other experts shared insights on addressing various corporate production scenarios. They discussed the selection and application of customized experience extraction plugins to convert employees' knowledge into reusable AI tools for organizations, which can stimulate further reflection.

The '2023 LLM Technology Exchange Conference' not only offered invaluable insights from industry experts but also attracted a significant number of professionals actively seeking solutions. The diverse participation enriched the event's content and fostered interdisciplinary exchanges. Essentially, beyond aligning LLM development with Taiwanese usage patterns, there exists an expectation that enterprises are capable of building a financially viable knowledge base while striking a delicate balance between information security and intelligence, laying a solid foundation for the future advancement and utilization of LLMs. For further details, please visit myLLM.tw.