



**AI 輕量化安全防護效能優化
與評測資料集擴充
專業勞務採購規範書
(A652001218)**

**工業技術研究院服科系統科技中心
2026/10/28**



目錄

一、	專案名稱	3
二、	專案目標	3
三、	委託規劃說明	3
四、	專案時程與交付文件.....	4
五、	結案驗收與付款方式.....	4



一、 專案名稱

「AI 輕量化安全防護效能優化與評測資料集擴充」

二、 專案目標

1. 強化輕量化 AI 安全防護模型之安全辨識效能，評估不同模型於中英文 Prompt 之辨識能力及邊緣運算環境之適用性。
2. 擴充 AI 安全評測資料集，涵蓋正常提示與惡意提示等情境，提升安全評測資料之多樣性與完整性。
3. 研析多模型分層串聯、信心分流及協同防護等技術策略，提出兼顧中文安全辨識能力與運算資源效率之優化方案。

三、 專案規劃說明

隨著人工智慧（Artificial Intelligence，AI）與邊緣 AI（Edge AI）技術快速發展，AI 模型逐步應用於物聯網、車載系統及智慧監控等終端裝置。然而，AI 模型在實際應用過程中，可能面臨提示注入、越獄攻擊及對抗性輸入等安全威脅，傳統資訊安全機制未必能有效辨識此類風險，使 AI 模型之安全防護能力成為重要技術課題。

考量邊緣運算裝置之運算能力、記憶體資源及即時處理需求，AI 安全防護機制須兼顧安全辨識能力與運算效率。輕量化安全防護模型雖具有較低的資源需求，惟不同模型於惡意輸入辨識能力、推論時間及記憶體使用方面各具特性，尤其面對中文提示及多樣化攻擊情境時，仍可能存在辨識準確性不足及誤判、漏判等問題。

此外，AI 安全防護模型之評測成效亦受到資料集規模、語言分布、攻擊類型及標記方式等因素影響。為提升評測結果之完整性與代表性，有必要擴充國內外公開安全評測資料來源，涵蓋不同語言及安全風險情境，並針對資料集特性、品質與適用性進行分析，以建立



更完善之安全防護效能評估基礎。

爰此，本案規劃委託辦理輕量化 AI 安全防護效能分析與評測資料集擴充研究，並針對中文安全辨識能力及多模型協同防護需求，研析分層式模型串聯、信心分流等技術策略之適用性與可行性，提出兼顧安全防護能力及運算資源效率之優化建議。

四、 專案時程與交付文件

1. 本專案期程為自採購生效日起至 114 年 11 月 13 日。
2. 於 115 年 11 月 15 日(含)前，完成「AI 輕量化安全防護效能優化與評測資料集擴充成果報告」電子檔一份

五、 結案驗收與付款方式

1. 廠商應依據本委託規定之工作時程及完成期限，完成相關工作後將應交付之文件送交本中心審查及確認
2. 期末成果經驗收合格後付款 100%