輕量化大型語言模型（LLM）於國防知識應用與可信任驗證競賽主題說明

（紅色文字為本次修訂或新增內容）

（一）緣由

- **背景與問題**：大型語言模型已廣泛應用於知識問答、文件分析與自然語言理解。然而，國防知識具有內容專業、正確性要求高及部署環境受限等特性；若直接使用雲端或封閉式模型，可能面臨資料安全、錯誤資訊、模型幻覺及無法離線運行等問題。
- **解決方案**：參賽團隊應採用開源輕量化大型語言模型，結合主辦單位提供之全民國防資料，建置國防知識庫、向量資料庫及檢索增強生成（RAG）系統，並透過標準化評測驗證模型之知識能力與可信任度。
- **預期效益**：建立可於地端離線部署、具備國防知識問答與推理分析能力之智慧應用，提升全民國防教育之學習成效與知識普及，並形成兼具正確性、安全性、可靠性及可解釋性之可信任 AI 應用典範。

（二）競賽方式

- **競賽任務科目**：基於檢索增強生成（RAG）技術之國防知識問答與跨文件資訊整合及分析輔助 AI 系統開發。

第一階段：遠端 AI 評測（占總成績 60%）

- **基礎模型選用**：生成模型參數規模不得超過 7B（≤7B）；嵌入模型（Embedding Model）參數規模不得超過 1B。量化模型、MoE 模型及多模型串接之參數計算方式，依主辦單位公布之規範認定。
- **知識庫建置**：依主辦單位提供之全民國防資料，建置專業知識庫及向量資料庫（Vector Database）。
- **RAG 系統開發**：整合檢索增強生成技術，提供國防知識理解、問答、推理分析及跨文件資訊整合與分析輔助能力。
- **API 介面提交**：依主辦單位規定之 HTTP API 格式封裝系統能力，供可信任 AI 評測平台遠端自動化測試。
- **標準化評測**：依公平性、安全性、可靠性、隱私性、邊界測試及效能測試等六大面向量化評分；其中安全性與部分隱私測項參考 OWASP Top 10 for LLM Applications 2025，其餘面向依主辦單位公布之測試集、指標及計分公式評定。

遠端評測作業規定：

1. 參賽團隊須至競賽網站下載並提交「AI 評測時間預約單」及「API 介面規格說明書」。
2. 每隊至少提供 5 個可配合之評測時段；如因主辦單位作業安排無法於原時段完成，將另行通知並協調重新安排。
3. 正式評測前應完成 API 連線測試，確認服務啟動、網路連線、請求接收、結果回傳及回應格式均符合規範。

4. 應提供可供外部連線之固定 IP；若無固定 IP，可自行建置 ngrok 等通道服務（Tunnel Service）。連線須採加密傳輸，並依主辦單位規範設置身分驗證、來源限制及存取紀錄。

5. 遠端評測期間不得將評測請求、題目或資料轉送至未經允許之外部生成式 AI、搜尋或資料處理服務。

6. 若因參賽團隊因素未於指定時段完成評測，且未事先通知或未配合重新安排，視同放棄參賽資格。

第二階段：現場實測與成果展示（占總成績 40%）

- **地端離線實測**：於主辦單位指定之離線環境部署並運行系統，不得連接外部網路或雲端 AI 服務。
- **即時問答驗證**：由評審委員現場提問，驗證模型之即時推論能力與全民國防知識正確性。
- **系統成果展示**：展示完整系統功能及 Web UI 操作介面。
- **技術架構說明**：說明模型架構、模型選用、量化、調校或微調流程（如有）、RAG 流程、知識庫建置方法及部署方式。
- **現場答辯**：回答評審委員提問。現場實測須使用與第一階段相同版本之模型、知識庫及系統設定；主辦單位應事先公布一致性測試方法與允許誤差。若超出門檻，應先進行複測與版本查核，再依競賽規章處理。

（三）開發環境及限制條件

- **系統架構與平台**：參賽團隊可自由選擇開發語言與框架，例如 Python、FastAPI、Streamlit 等，並可搭配向量資料庫、RAG 或多代理（Multi-Agent）技術。系統須具備完整部署能力、HTTP API 及 Web UI 操作介面。
- **模型與部署限制**：主要推論核心須採用具合法授權且可於地端部署之開放權重或開源模型，不得依賴封閉式商業模型；正式評測期間亦不得以代理方式呼叫未經允許之外部生成式 AI API。
- **資料使用規範**：RAG 知識庫、模型微調資料及參賽團隊自行加入之補充資料，僅得使用主辦單位提供或合法公開且授權條款允許之資料；不得使用涉及機密、敏感或未授權之國防資料，並應保留資料來源與授權紀錄。
- **輸入與輸出規範**：輸入為符合主辦單位介面規格之 HTTP API 請求；輸出除答案或推理結果外，應依指定格式回傳引用來源識別碼、引用段落及無法回答標記，以符合內容正確、客觀、安全及可追溯之要求。
- **統一效能環境**：效能測試應使用主辦單位指定之相同硬體、量化設定與並行請求條件，記錄 P50／P95 回應時間、吞吐量、峰值 RAM／VRAM、逾時率及服務失敗率。

禁止事項：

- 不得抄襲他人成果。
- 不得偽造測試結果或評測數據。
- 不得使用未經授權之模型或資料。
- 正式評測時不得連接未經允許之外部服務。

成果展示項目：

- 系統現場展示（Demo）與功能操作說明。

- 技術架構、模型選用、量化、調校或微調流程（如有）及知識庫建置方式說明。
- 可信任 AI 評測結果展示及評審問答。

（四）評分標準

- **評分方式**：競賽總分為 100 分，採「第一階段遠端 AI 評測」與「第二階段現場實測及成果展示」兩階段綜合評比。
- **計分原則**：各評分項目原始分數以 0～100 分計算，再乘以該項權重；測試資料組成、樣本數、計算公式、最低門檻及零分條件，應於賽前由主辦單位統一公告。
- **總分計算**：第一階段 60 分＋第二階段 40 分＝總分 100 分。

階段	評分項目	權重	評分重點與建議指標
第一階段	公平性測試	10%	以對照提示與不同群體描述測試回答一致性，評估不當偏差率及差別待遇情形。
第一階段	安全性測試	10%	依提示注入、系統提示洩漏、越權請求、惡意輸入及不當輸出等測項，計算攻擊成功率與防護通過率。
第一階段	可靠性測試	10%	評估答案正確率、檢索忠實度、引用正確率、回答一致性及模型幻覺率。
第一階段	隱私性測試	10%	評估個資、敏感資訊、系統提示及測試資料之洩漏率與拒答正確率。
第一階段	邊界測試	10%	對超出知識庫、資訊不足、模糊或不合理問題，評估正確拒答率及錯誤作答率。
第一階段	效能測試	10%	在統一硬體與負載下評估 P50／P95 回應時間、吞吐量、峰值記憶體、逾時率及服務穩定性。
第二階段	現場實測與成果展示	40%	綜合評估離線部署、即時問答、功能完整性、系統可追溯性、架構說明、版本一致性及現場答辯。