附件一

新北市政府人工智慧應用風險評估檢核表

填寫目的：參酌人工智慧基本法規定及人工智慧風險分類框架模式，評估並掌握本府各機關人工智慧系統與服務之潛在風險。

填寫範圍：本府各機關自行開發、委外建置或採購之各類人工智慧系統與服務，例如生成式人工智慧、決策輔助與預測分析系統、機器學習模型或其他具自動化決策能力之人工智慧應用等。

填寫說明：本表由人工智慧系統承辦人或機關指定人員填寫，作為符合人工智慧基本法第十九條進行風險評估與訂定內控管理措施之基礎。

判定說明：人工智慧系統與服務經綜合判定結果應依核心風險識別為之；有一定程度風險情形者，應依風險評估檢核表所列之嚴重程度及發生機率，判定其風險類別。

本要點所稱重大功能變更，指對人工智慧系統與服務本身，或其使用條件、情境或範圍所進行之更新，致改變其功能、底層架構或效能，而足以實質影響其應用結果或風險程度，並使先前之評估相關文件可能失其適用性或產生誤導之虞者，包括下列各款情形之一：

1. 模型核心邏輯或演算法架構變更。
2. 訓練資料來源、範疇或標註方式發生重大異動（如異動比例達百分之五十，或資料性質顯著改變）。
3. 系統適用對象、利害關係人、使用情境或決策範疇擴大或變更。
4. 輸出結果之後處理邏輯或門檻設定變更，致影響最終決策或裁量結果。
5. 其他經評估認定足以影響系統風險等級之變更。

一、應用情境盤點

本表格目的為掌握人工智慧系統與服務之基本資訊。

檢核項目	填寫內容 （評估結果）	依據來源	備註 參考文件
1.1 系統與服務名稱	[請填寫]	人工智慧風險分類框架(附錄1)	系統清冊
1.2 系統與服務具體用途與服務對象	【具體用途】 [請填寫] 【服務對象】 □政府內部使用 □對外公開服務	人工智慧風險分類框架(附錄1)	系統清冊
1.3 核心人工智慧技術與開發方式			
1.3.1 生成式AI應用（如大型語言模型LLM）	□自行建置 □採購 □未使用	人工智慧風險分類框架(附錄1)	採購或系統文件
1.3.2 鑑別式AI應用	□自行建置 □採購 □未使用		
1.3.3 其他人工智慧技術	□自行建置： _____ □採購： _____ □未使用		

1.4 辨識利害關係人			
1.4.1 負責開發人工智慧系統與服務	□自行開發 □委外開發 □僅進行採購（類SaaS 服務）	人工智慧風險分類框架(附錄1)	業務作業說明
1.4.2 負責部署人工智慧系統與服務	□自行部署 □委外部署 □僅進行採購（類SaaS 服務）		
1.4.3 可能因人工智慧系統與服務影響之利害關係人 *可複選	□民眾 □本機關公務員 □跨機關公務員 □其他：————		

二、核心風險識別與危害評估

本部分參考人工智慧風險分類框架，用以檢視並確認人工智慧系統與服務是否存在風險。填寫說明如下：

(一)若 2.0 任一項勾選「是」，該項風險應依本府資通安全維護計畫或個人資料相關法令辦理，其結果不納入後續 2.4 嚴重程度、2.5 風險發生機率判定及風險矩陣計算。

(二)若檢核項目 2.1、2.2、2.3 任一項勾選「是」，請接續填寫 2.4 嚴重程度判定及 2.5 發生機率。

(三)若檢核項目 2.1、2.2、2.3 均勾選「否」，則就風險類型一至三判定為「無顯著風險」，免進行後續嚴重程度判定、發生機率及風險矩陣分類。

檢核項目	評估結果	情境簡述	依據來源
2.0 風險類型零：資安及個資問題 本人工智慧系統與服務、訓練資料、輸入指令、輸出內容，是否可能因其業務流程、使用情境或設計邏輯，而有違反資通安全或個資保護之疑慮？	資通安全疑慮： □是 □否 個資保護疑慮： □是 □否	[若評估結果為是，請簡述]	
2.1 風險類型一：人工智慧系統與服務本身之技術設計缺陷 本人工智慧系統與服務，是否可能因其業務流程、使用情境或設計邏輯，而發生演算法偏見、產生錯誤或誤導訊息、缺乏透明度或存在安全漏洞等情況，導致對特定群體不公平或決策錯誤？	□是 □否	[若評估結果為是，請簡述]	人工智慧風險分類框架（二、識別風險）
2.2. 風險類型二：部署後操作及人機互動問題			
2.2.1 本人工智慧系統與服務，是否可能因其業務流程、使用情境或設計邏輯，而導致公務員過度依賴而喪失獨立判斷能力？	□是 □否	[若評估結果為是，請簡述]	人工智慧風險分類框架（二、識別風險）
2.2.2 本人工智慧系統與服務，是否可能因其業務流程、使用情境或設計邏輯，而導致人工智慧系統與服務遭濫用以生成違法內容、深偽詐欺？	□是 □否	[若評估結果為是，請簡述]	

2.2.3 本人工智慧系統與服務，是否可能因其業務流程、使用情境或設計邏輯，而導致人工智慧系統與服務遭利用於網路攻擊？	□是 □否	[若評估結果為是，請簡述]	
2.3 風險類型三：社會結構與環境衝擊			
2.3.1 本人工智慧系統與服務，是否可能因其業務流程、使用情境或設計邏輯之應用，而加劇社會不平等之疑慮或風險？	□是 □否	[若評估結果為是，請簡述]	人工智慧風險分類框架（二、識別風險）
2.3.2 本人工智慧系統與服務，是否可能因其業務流程、使用情境或設計邏輯之應用，而耗費能源造成環境傷害？	□是 □否	[若評估結果為是，請簡述]	
2.3.3 本人工智慧系統與服務，是否可能因其業務流程、使用情境或設計邏輯之應用，而遭利用於認知作戰，影響公眾信任？	□是 □否	[若評估結果為是，請簡述]	
2.4 嚴重程度判定 若檢核項目2.1、2.2、2.3任一項勾選「是」，請接續進行嚴重程度判定，並採「擇一最高等級」單選勾選（例如：同時符合嚴重與非常嚴重之指標時，請勾選非常嚴重）；若經檢視，本案情境若均未達下方「非常嚴重」、「嚴重」及「一般」之指標門檻者，則下列各項請均勾選「否」。			
嚴重程度(依風險發生時之危害程度判定)		**評估結果**	**依據來源**
2.4.1 **【非常嚴重】** 符合下列情形之一者屬之： 1. 上開風險發生，可能對機關之營運、資產或信譽等方面產生非常嚴重或災難性之影響。		□是 □否	參考人工智慧基本法、「資通安全責任等

2. 上開風險發生，可能影響他人合法權益或機關執行業務之公正性及正當性，並使機關所屬人員負刑事責任。 3. 上開風險發生，可能導致特定群體（如特定性別、年齡、身心障礙或弱勢族群）在政府資源分配、資格審查或行政決策上遭受嚴重不公平對待。 4. 上開風險發生，可能違反兒少最佳利益原則，或對其身心發展造成嚴重負面影響。 5. 上開風險發生，可能引發社會恐慌，或造成對人類基本權利、生命安全、財產保障、社會秩序或政府公信力之嚴重危害。		級分級辦法」之「附表九、資通系統防護需求分級原則」
2.4.2 【嚴重】 符合下列情形之一者屬之： 1. 上開風險發生，可能對機關之營運、資產或信譽等方面產生嚴重之影響。 2. 上開風險發生，可能影響他人合法權益或機關執行業務之公正性及正當性，並使機關或其所屬人員受行政罰、懲戒或懲處。 3. 上開風險發生，可能導致特定群體（如特定性別、年齡、身心障礙或弱勢族群）在政府資源分配、資格審查或行政決策上遭受不公平對待。 4. 上開風險發生，可能違反兒少最佳利益原則，或對其身心發展造成負面影響。 5. 上開風險發生，可能引發社會恐慌，或造成對人類基本權利、生命安全、財產保障、社會秩序或政府公信力之危害。	□是 □否	參考人工智慧基本法、「資通安全責任等級分級辦法」之「附表九、資通系統防護需求分級原則」

2.4.3 **【一般】** 符合下列情形之一者屬之： 1. 上開風險發生，可能對機關之營運、資產或信譽等方面產生有限之影響。 2. 上開風險發生，可能發生違反相關規範之情形。	□是 □否	參考人工智慧基本法、「資通安全責任等級分級辦法」之「附表九、資通系統防護需求分級原則」
2.5 風險發生機率 若檢核項目2.1、2.2、2.3任一項勾選「是」，且已判定嚴重程度者，請接續進行風險發生機率判定。 發生機率為參考區間，機率判定得參考歷史發生紀錄、系統特性、使用情境及其他相關風險評估資料綜合判斷之。 機率判定應以未施加額外控制措施時之固有發生可能性為準，不得以既有防護措施降低機率評等。		
發生機率	□高：指發生機率約在百分之六十以上，且於系統與服務運行期間內極可能發生。 □中：指發生機率約在百分之三十以上未達百分之六十，且於系統與服務運行期間內可能發生。 □低：指發生機率未達百分之三十，於系統與服務運行期間內不太可能發生，或僅於特殊或極端情境下始可能發生。	參考澳洲數位轉型署，人工智慧衝擊影響評估工具（AI Impact Assessment Tool）

三、人工智慧內控管理與應對機制（Internal Controls & Response）

本府各機關請依據前述「核心風險識別與危害評估」之結果（含項目2.0、2.1至2.3核心風險、2.4嚴重程度及2.5發生機率），依下列規定辦理後續內控管理與措施填列：

1. 資安與個資風險處理：若2.0任何一項勾選「是」，則3.6項為必辦項目，且該項風險應依本府資通安全維護計畫或個人資料相關法令處理，本府資訊中心得視內部控管或查核需要，要求本府各機關提供相關佐證文件。（例如資通安全維護計畫、風險評鑑報告、風險評鑑彙整表、風險改善計畫、個人資料風險評估紀錄或其他足資證明之文件）。

   本案個資與資安風險識別：

   □有資安風險之虞□有個資風險之虞□無上述風險

2. 人工智慧風險等級判定

   本府各機關請依據前述核心風險評估結果，確認風險類別：

   （1）若檢核項目2.1至2.3均為「否」，或2.4嚴重程度均為「否」，則風險類型一至三屬「無顯著風險」，免予對照下方風險矩陣，並逕於下方勾選綜合判定結果。

   （2）若檢核項目2.1至2.3任一項勾選「是」，並已完成2.4嚴重程度及2.5發生機率判定者，請依其判定結果對照下方風險矩陣劃分歸屬類別（甲類、乙類、丙類）：

發生機率 / 嚴重程度	高	中	低
非常嚴重	甲	甲	乙
嚴重	甲	乙	丙
一般	乙	丙	丙

本案風險綜合判定結果（請擇一勾選）：

□甲類□乙類□丙類□無顯著風險

3. 內控措施落實填列說明

本府各機關請依據前項「本案風險綜合判定結果」，依下列規定填列後續內控措施表：

（1）判定為甲類或乙類者：3.1至3.6項為必辦項目，本府各機關應逐項採取控制措施並填列辦理情形（乙類之落實優先順序得次於甲類）。

（2）判定為丙類者：3.1及3.2項為必辦項目，應採取控制措施並填列辦理情形；其餘3.3至3.6項仍應於表單進行評估填列，惟得免採取該等內控措施。

（3）判定為「無顯著風險」者：得免辦理第3.1至第3.6項內控措施，惟仍應依附表填列評估結果，並保留相關紀錄備查。

（4）前述各類別之必辦項目，若系統與服務性質確實無涉決策（如3.1）或無涉內容產出（如3.2）等特殊情境者，得免採取措施，惟仍應於表單勾選「不適用」並具體敘明理由。

4. 人工智慧內控措施落實表

檢核項目	落實結果	辦理/不適用說明	參考來源
3.1 人類監督保留：針對人工智慧提供之決策，作業流程中是否設有適當之人為介入與最終覆核機制，避免完全自動裁決？	□是 □否 □不適用	[請填寫]	人工智慧基本法第 4 條第 2 款

3.2 **透明標示**：系統與服務之產出是否進行適當資訊揭露(如標註「由人工智慧生成」)?	□是 □否 □不適用	[請填寫]	人工智慧基本法第 5 條
3.3 **注意事項或警語**：本系統與服務是否依其應用特性與潛在風險，於服務介面或相關文件中明確標示注意事項或使用警語？	□是 □否 □不適用	[請填寫]	人工智慧基本法第 5 條
3.4 **建立救濟機制**：針對受人工智慧輔助決策影響之民眾，是否提供明確之申訴管道、決策解釋與權益救濟機制？	□是 □否 □不適用	[請填寫]	人工智慧基本法第 17 條
3.5 **強化防護**：是否於採購契約或系統與服務規格中，明確要求供應商落實資安及資料治理相關機制？	□是 □否 □不適用	[請填寫]	人工智慧基本法第 4 條第 3款、第 4 款
3.6 **上線前檢測**：於系統上線前應提供AI檢測申請單（如附件二）送至本府資訊中心進行檢測。	□是 □否 □不適用	[請填寫]	

系統承辦人：＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿(簽章)

機關首長或其授權人員：＿＿＿＿＿＿＿＿＿＿＿＿ (簽章)