新北市政府教育局使用人工智慧補充規定

1、 **依據**：參考行政院2023年《行政院及所屬機關（構）使用生成式AI參考指引》、數位發展部2024年《公部門人工智慧應用參考手冊》與本府2025年《新北市政府使用人工智慧作業指引》訂定。

2、 **目的**：為引導新北市政府教育局(以下簡稱本局)所屬人員，能以負責任、合倫理、安全且有效率的方式使用人工智慧（AI）技術，特訂定本補充規定。

3、 **適用範圍**：本補充規定適用於本局及機關學校所屬人員，於執行公務時，所使用或採購之AI系統，包含生成式AI服務、分析型(鑑別式)AI系統及第三方AI服務。

4、 **分級說明**：本局將AI的可應用程度劃分為五個等級，從「不可應用」至「完全可應用」，說明如下：

級別(Level)	簡要原則(Principle)	管理規範(對應風險等級)
Level 1－**不可應用**	**禁止使用**：涉及高度敏感性(如健康、基因、生活)、法律或倫理風險，可能造成嚴重誤導、歧視或權益侵害。	**嚴格禁止**。違反《行政院及所屬機關（構）使用生成式AI參考指引》第三點規定。
Level 2－**嚴格限制**	**高度限制**：原則上不使用，僅在特殊情況下於嚴密管控下試行。需經主管批准並採取額外防護措施。	**強制性實質人工審查**。部署前必須完成正式之「資料保護與倫理衝擊評估」（DPIA）。
Level 3－**審慎應用**	**有條件使用**：可在有限範圍內使用，但須謹慎評估風險並保留人工監督。AI僅作為輔助工具，產出需經人工審查確認正確性與適法性。	**事實查核義務**。承辦人須負擔最終之事實查核與潤飾責任。嚴禁將「密」等級以上公文或敏感個資輸入。
Level 4－**廣泛應用**	**原則允許**：可大範圍應用於日常業務情境，視為安全可靠。應遵循本規範的一般準則，例如定期評估模型表現。	**遵守資安規範**。依循本局既有之資訊安全及文書處理規範使用。

※注意：本分級應視個案特性彈性調整。人員在評估AI導入時，若有不確定之處，可參考附件自我檢核表，提請單位內資科承辦或倫理審查小組協助評估。

5、 合作說明

(1) **數據授權**：提供給廠商用於模型開發或服務之任何資料，應明確約定其使用範圍與目的。廠商僅能在授權範圍內使用該等資料，不可將資料用於未經允許的其他用途，並不得轉授權第三方。

(2) **訓練資料歸屬**：原始資料的所有權應歸屬原單位。若廠商利用公務資料訓練AI模型，應約定本局對該模型或產出成果擁有適當的權利或共同權益。此舉旨在防止廠商於合約範圍外自行保留或利用以單位資料訓練之模型。

(3) **安全責任**：合作雙方須明確各自承擔的資訊安全責任。廠商應遵循國家資安法規及本局資安政策，妥善防護所處理的所有資料。合約中應要求廠商採取必要的安全措施（如資料加密、存取控制、異常監控），並約定一旦發生資料外洩或安全事故，廠商須立即通報並承擔相應責任。

(4) **著作權/智慧財產權合規**：使用AI產生內容時須確保不侵犯第三方著作權。遵守版權及人格權相關法規，以補強法規遵循面向的完整性。

(5) **合作終止與資料刪除**：在合作關係終止時，廠商應立即歸還或銷毀所有由原單位提供的資料及經由該等資料訓練所得的模型或分析結果。廠商須提供書面證明完成資料刪除，並確保不得於合作終止後繼續保留任何本局機密資訊或個人資料。

6、 **補充規定修正**：本補充規定經「新北市資訊科技諮詢委員會」討論通過，未來隨著技術與法規的演進，本局將定期檢視內容，確保教育AI應用持續符合最新的安全標準與社會期許。

新北市教育局所屬各機關AI應用自我檢核表

1、 應用情境基本資訊

項目	說明與填寫內容	備註
1. 應用名稱	（請提供此AI應用的名稱或專案代碼。）	
2. 應用目的	請簡要說明此AI應用旨在解決的核心問題或痛點，並說明預期達成的效益（例如：減少工作時間、提升工作效率或協助做出更好的判斷）。	
3. 業務範疇	此應用主要屬於哪類行政業務？（可複選）	
	☐ 文書與庶務（公文草擬、會議支援）	
	☐ 數據分析與決策（教育統計、資源配置輔助）	
	☐ 公眾服務（智能客服、資訊推播）	
	☐ 資源與人事管理（採購文件輔助、研習管理）	
	☐ 政策規劃或教育訓練	
	☐ 其他應用：	
4. AI 產出物類型	AI系統的最終輸出是什麼？	
	☐ 內部建議或洞察（Insight） ☐ 行政文書初稿或摘要 ☐ 決策依據（作為人工決策的**主要參考**） ☐ 直接提供給外部使用者（如智能客服回覆） ☐ 其他：	

2、 風險等級與人工監督評估(本段請視需求應用)

項目	評估內容	符合請勾選（或填寫）	備註
1. 權益影響評估	此AI應用是否涉及**直接影響**學生或教職員的**重大權益、安全、教育或職業機會**？（例如：學籍異動、獎懲建議、資源分配、績效評估）。	☐ 是（若為「是」，則至少為 Level 2：高風險） ☐ 否	
2. 機密性評估	此AI應用是否需處理**機密文書、個人身分資料**或**高度敏感性資料**？	☐ 是（若為「是」，則**嚴禁**使用外部AI平台） ☐ 否	
3. 風險級別初判	根據應用目的及權益影響程度，此情境最可能屬於哪一級別？		
	☐ Level 1 – **不可應用（嚴格禁止）**：例如AI**直接決定**學生處分或自動化品行評分。		
	☐ Level 2 – **嚴格限制（高風險）**：例如AI作為**主要依據**分配校際經費；需**實質人工審查**。		
	☐ Level 3 – **審慎應用（中/低風險）**：例如草擬行政文件初稿；AI作為**輔助工具**。		
	☐ Level 4 – **廣泛（最小風險）**：例如拼字檢查、自動排程。		
4. 人類參與決策	無論風險等級為何，最終決策者（業務承辦人）應對AI產出負起**最終責任**,。請說明**誰**將進行最終的審查、校正與決行。	相關單位：	
5. 稽核紀錄	若屬於 Level 2（高風險）應用，是否規劃**詳實記錄**AI的演算建議與最終人工決策，以備查核？	☐ 已規劃實施稽核紀錄 ☐ 待規劃相關稽核文件 ☐ 尚未規劃，原因：	

3、 AI倫理原則檢核

倫理原則	具體檢核項目	符合請勾選	備註
問責性（Accountability）	1. 是否已確保AI產出內容將經**人工專業判斷與事實查核**，避免完全依賴AI,？	□ 是 □ 否	
數據隱私（Data Privacy）	2. 是否已規劃**嚴禁**將機敏個資（如學籍、健康紀錄）上傳至**未經核可之外部AI平台**？	□ 是 □ 否	
	3. 是否遵循「**資料最小化**」原則，僅使用達成目標所需之最少資料,？	□ 是 □ 否	
公平性（Fairness）	4. 若用於資源分配或篩選，是否規劃進行**偏見稽核**（Bias Audit），以審視結果是否對特定群體造成系統性不利？	□ 是 □ 否	
透明度（Transparency）	5. 若作為公眾服務（如智能客服），是否**清楚標示為AI服務**，並提供轉接真人服務選項？	□ 是 □ 否	
資訊安全（Security）	6. 是否已完成AI資安意識培訓，警覺**提示注入攻擊**及**深度偽造**等新型資安威脅？	□ 是 □ 否	
	7. 若涉及**外部廠商**，合約中是否已明確約定**數據授權範圍**與**安全責任**，並要求**資料加密**？	□ 是 □ 否	若適用

4、 實施與數據評估(本段請視需求應用)

項目	評估內容	說明與現況（請簡述）	備註
1. 資料量/品質	**資料是否足夠？** 是否已評估訓練資料的**完整性、準確性、真實性與即時性**？		
2. 資料標記需求	AI模型訓練是否需要手動**標記資料**？（若需要，請預留足夠前置時間）。		
3. AI 導入模式	專案計畫採取哪種導入模式？ ☐ 採購現成商業服務模組 ☐ 外部廠商客製化建置 ☐ 內部自行建置 ☐ 其他說明：		
4. 專業領域解讀	誰將負責**解讀AI分析的結果**，特別是當AI找出「關聯性」不等於「因果性」**時，由哪位**領域專家或規劃師進行詮釋與校正?		
5. 潛在限制	是否已預先評估AI輸出的**不確定性**或「幻覺」現象，並規劃配套措施？		
6. 營運管理	是否已規劃模型部署後的**持續監控機制**，以追蹤模型效能與潛在的**資料飄移**問題？		

5、 審核結論與建議

項目	審核結果	建議與應採取的額外措施（若有）
審核結論	□ 建議導入，風險可控（Level 3/4/5）	
	□ 嚴格限制，須進行額外評估（Level 2）：需完成正式之「資料保護與倫理衝擊評估」（DPIA）。	
	□ 不建議導入，應尋求替代方案（Level 1）。	
	審查單位：	

附件：分級範例簡要說明

級別(Level)	參考範例
Level 1 - **不可應用**	例如以AI資料審定**教師資格**、判定學生**處分**或自動化**品行評分**，遠端**生物辨識**用以進行紀律管理，皆可能受到資料偏見導致嚴重錯誤。
Level 2 - **嚴格限制**	進行校際**經費、補助款**或教師**員額**之分配。**判斷學生**之學籍異動、獎懲建議。或運用AI分析**去識別化**的敏感資訊。
Level 3 - **審慎應用**	運用AI草擬行政文件或報告**初稿**。使用AI自動**摘要**大量會議資料。智能客服提供法規、流程之自動**問答**。需注意資料庫、RAG效果、模型是否能正確提供。
Level 4 - **廣泛應用**	使用AI**翻譯**公開文件資料、將市民來信依內容自動**分類**、使用AI自動**排程與通知**（例如會議時間安排最佳化、例行通知發送）、文書處理軟體內建之**拼字檢查**、**文法建議**等通用輔助功能。