中華人民共和國香港特別行政區政府
數字政策辦公室

香港生成式人工智能研發中心
Hong Kong Generative AI | Research & Development Center

香港
生成式人工智能技術及應用指引

2025年4月

前言

香港特別行政區政府（特區政府）數字政策辦公室（數字辦）委託了透過「InnoHK 創新香港研發平台」撥款支持成立的香港生成式人工智能研發中心協助研究制定《香港生成式人工智能技術及應用指引》，以期就生成式人工智能技術及應用方面提供相關的守則及指引，供各界參考。

本指引闡述生成式人工智能的技術背景和治理原則，以及為技術開發者、服務提供者和服務使用者提供實用指南。

數字辦會持續審視生成式人工智能相關的最新技術及應用發展，定期更新指引的內容。通過技術開發者、服務提供者和服務使用者等廣泛持份者的共同努力，讓生成式人工智能更好地服務香港社會，為市民和業界帶來更多便利與福祉。

註：如本文件中英文版本內容存在歧義，請以英文版本為準，並通知數字政策辦公室

目錄

目錄

背景

生成式人工智能是以機器學習為代表的現代人工智能的重要分支。生成式人工智能借助各類機器學習算法，能夠根據人類複雜意圖和指令，自動生成文本、圖像、音頻、視頻等內容。近年來，隨著生成式人工智能技術取得突破性進展，相關產品及服務得到迅速應用和推廣。與其他人工智能應用及傳統互聯網應用相比，生成式人工智能的最大優勢在於具備更加個人化及便捷的內容生成能力，是人工智能邁向通用化進程的重要里程碑。

當前，生成式人工智能技術的變革影響全球，未來將在更多領域得到廣泛應用，對社會經濟發展及人類文明進程產生深遠影響。與此同時，該技術亦面臨難以預測的安全風險和倫理問題等挑戰，人工智能治理已成為世界各國共同面對的課題。需要確保在安全的範圍內發展和應用生成式人工智能技術。

香港特別行政區政府（特區政府）認識到，香港必須緊跟技術發展趨勢，制定與之適配的人工智能治理方案，積極推動人工智能技術發展，為香港經濟和社會的建設貢獻力量。為此，特區政府積極借鑒國際最佳實踐經驗，與本地產業及創新科技領域的專家們攜手合作，致力於提升香港人工智能行業生態的韌性。特區政府委託專注於生成式人工智能技術的香港生成式人工智能研發中心（HKGAI），針對開發及應用生成式人工智能技術，研究關於可信度、責任、倫理安全方面的指導原則，並提出相關建議。

受特區政府的委託，HKGAI 參考業界專家意見以及國際最佳實踐經驗，編制《香港生成式人工智能技術及應用指引》（指引）。指引旨在推動香港特別行政區內的相關持份者，在嚴格遵循科技倫理、道德準則以及法律規範的基礎上，安全且負責任地開展涉及生成式人工智能的相關業務及活動。同時，為各持份者提供切實可行的指引，有效協助應對生成式人工智能技術引發的安全問題及社會風險。

指引旨在讓香港生成式人工智能的發展能平衡創新和社會責任，確保技術發展的效益最大化，風險最小化，為香港開創一個強有力的生成式人工智能治理框架，推動各產業領域生成式人工智能的開放應用和健康發展。

指引特別針對以下持份者制定：

1. 技術開發者、委託他人開發技術或決定技術用途的人士（技術開發者）；
2. 服務提供者、平台提供者、利用現成技術二次提供附加功能和工具的人士（服務提供者）；
3. 服務使用者、生成式內容創作者和傳播者（服務使用者）。

隨著生成式人工智能技術日趨普及，技術應用面臨諸多安全和社會風險。指引從實踐角度出發，提供具體指導和建議。對於技術開發者，需關注資料洩露、模型偏見和錯誤等技術風險，指引將提供安全開發和設計的最佳實踐，確保技術的可靠和穩定性；服務提供者需關注服務的合規性和安全性，尤其是處理服務使用者資料和生成內容時，指引將提供建議，確保服務安全運行；服務使用者則需了解使用生成式人工智能服務的潛在風險，避免違例犯法及不道德行為，學習保護個人資料安全。指引將提供使用技術服務的建議，以及識別和應對潛在風險的措施。

1. 生成式人工智能簡述

1.1 生成式人工智能的概述

生成式人工智能是指利用各種機器學習算法，讓電腦系統能夠根據大量數據，以及複雜的人類意圖和指令，自動生成內容資訊，例如文字、圖像、音頻、視頻、程式碼或其他媒體。

與其他人工智能應用以及傳統互聯網應用相比，生成式人工智能具備內容生成及創作能力。該技術所提供的並非互聯網上已存在的內容，而是模型算法基於服務使用者輸入指令重新創作生成的內容。服務使用者可根據自身需求，生成符合期望的內容或解決方案，從而獲得更加個人化及定制化的服務體驗。

生成式人工智能技術屬於以機器學習為代表的現代人工智能領域的一個分支，是支撐生成式人工智能服務的底層算法、模型及架構。生成式人工智能服務是指根據生成式人工智能技術開發的產品、解決方案及服務。

生成式人工智能技術研發至應用的生命週期中，涉及**技術開發者、服務提供者、服務使用者**這三類重要角色。

1.2 生成式人工智能技術的主要原理

生成式人工智能技術基於深度學習中的生成模型，通過無監督或自監督、自回歸等多種訓練策略，使模型學習大量訓練數據的分佈特徵，從而生成內容。生成的方法和策略因模型而異，可以是從隨機雜訊或初始輸入逐步生成，也可以通過壓縮數據形成隱空間，並在隱空間中採樣然後解碼生成數據。關鍵技術包括隱空間學習、概率建模、生成策略等，這些技術共同作用，使其能夠創造出多樣化的內容。生成式人工智能技術可以生成多模態內容，包括文本、代碼、圖像、音頻、視頻等。

目前主流模型包括生成對抗網絡（GANs）、變分自編碼器（VAEs）、自回歸模型、擴散模型（Diffusion Models）、Transformer 模型等：

- **GAN** 以相互競爭方式訓練兩個神經網絡，生成器網絡生成數據，判別器網絡判斷數據是否真實。通過對抗訓練不斷優化，從而使生成器能夠生成越來越真實的數據。
- **VAE** 使用編碼解碼的生成方法，將輸入數據映射到隱空間的分佈，通常是高斯分佈，然後從該分佈中採樣生成新的樣本。

- **自回歸模型**基於逐步生成的思想，將已生成的內容作爲上下文輸入，預測下一個元素從而生成新的數據，適用於生成有序的數據序列。
- **擴散模型**使用擴散去噪的生成方法，先逐步添加雜訊到原始數據，再通過學習逆過程來從雜訊中重構出新數據
- **Transformer 模型**基於自注意力機制和前饋神經網絡構建，採用編碼器-解碼器架構。生成過程主要依賴在用編碼器壓縮大量數據形成的語言隱空間中，通過解碼器利用自注意力機制處理上下文信息，並通過自回歸方式採樣，逐步生成新內容。

現時，「生成式人工智能」(Generative AI) 一詞主要指基於 Transformer 架構的大型基礎模型，包括大型語言模型 (LLMs) 以及視覺-語言模型 (VLMs)。這些模型採用自注意力機制 (self-attention) 和前饋神經網絡 (feed-forward neural networks)，並運行於編碼器-解碼器 (encoder-decoder) 架構之上。在生成過程中，模型會壓縮大量數據以形成隱空間，然後利用自注意力機制處理語境，並以自回歸方式生成新的內容。

基於 Transformer 的模型在生成多模態內容方面表現卓越，包括文字、程式碼、圖像、音頻和視頻。即使是通過從隨機圖樣中逐步去除噪聲來生成圖像的擴散模型(diffusion models)，通常亦會結合 VLMs 來處理文字提示 (textual prompts)。

雖然較早期的技術如生成對抗網絡 (GANs)、變分自編碼器 (VAEs) 及純自回歸模型 (pure autoregressive models) 對該領域作出了重要貢獻，但現代的生成式人工智能已主要由基於 Transformer 架構的模型所驅動，能夠以前所未有的能力跨多種模態進行處理與生成。

模型預訓練、微調和推理是生成式人工智能技術生命週期的重要階段，共同構成了從構建到部署的一套完整流程。

預訓練是指在大規模無標注或弱標注數據上對模型進行初步訓練的過程，目的是讓模型學習數據中的通用特徵和知識。這一階段通常採用無監督學習框架，例如自監督或自回歸方法。通過自回歸方式訓練的模型能夠有效學習數據分佈規律，從而具有強大的泛化能力。學習的知識可以輕鬆轉移到特定的任務上，從而減少對特定領域標注數據的需求。

微調是提升模型在特定任務或領域上性能表現的關鍵步驟。在模型預訓練完成後，模型需要在針對特定需求的小規模標注數據集進一步訓練，模型的部分層或全部層參數將會根據目標任務的數據進行微調，從而掌握特定領域知識。微調針對具體應用場景進行模型優化，相比從零開始訓練，在已有的預訓練模型上微調，能夠減少訓練時間和計算成本，大幅降低大規模參數模型的使用門檻。常見的微調技術包括指令微調、人類反饋強化學習 (RLHF)、直接偏好優化、低秩自適應 (LoRA) 等。

推理是生成式人工智能生命週期中的最後一個環節，也是模型實際應用的關鍵階段。在推理階段，完成訓練的模型將停止更新自身參數，轉而根據輸入的數據進行即時處理，以生成輸出。為了令訓練後的模型能處理服務使用者的輸入或實際應用中的數據，需要將其以服務的形式部署並上線至生產環境，從而滿足不同場景需求。推理的速度、效率和準確性是評估模型性能的關鍵指標，提升推理效率可以採用模型壓縮、優化等技術手段。

1.3 內容生成的主要模態

1.3.1 文本生成介紹

文本生成是指利用自然語言處理技術從輸入數據中自動生成自然語言文本，通常依賴於自然語言處理技術。目前最熱門的大語言模型也是文本生成的關鍵技術之一。模型通過從大量文本數據中學習語言模式和規律，能根據既定的輸入生成連貫且符合語法的文本。文本生成有兩種基本模式：

- 單模態文本生成在生成文本時，僅以文本作為輸入和輸出，根據既定的文本提示或上下文，預測並生成後續的文本內容。
- 多模態生成能夠同時處理多種類型的數據，如文本、圖像、音頻、視頻等，並在這些不同模態的數據之間建立聯繫和交互，從而生成文本。

1.3.2 圖像生成介紹

圖像生成是利用生成技術，通過對大量圖像數據的學習和分析，自動生成全新的圖像內容。通常將深度學習中的生成對抗網絡或擴散模型作為其訓練方法。圖像生成的模式有：

- 文生圖模式下，服務使用者輸入文字描述，模型解析語義信息並將其轉化為向量表示，基於學習到的知識在隱空間生成並解碼為圖像。
- 圖生成圖模式則是對輸入圖像進行特徵提取和編碼，在隱空間中操作變換後重構為新圖像。

1.3.3 音頻生成介紹

音頻生成是指利用人工智能技術，根據輸入的數據合成對應的聲音波形的過程，通常依賴於深度神經網絡來構建音頻信號的頻譜。音頻生成的模式為：

- 文生音頻將文字輸入轉化為音頻輸出，通過理解輸入的文本信息從而生成與文本內容相符或相關的音頻（如為輸入的歌詞譜曲）。

- 音頻生成音頻是指根據已有的音頻生成新的音頻，生成模型可以學習已有音頻樣本的規律和特徵，並基於這些規律和特徵生成相關新的音頻內容（如聽聲譜曲）。

1.3.4 視頻生成介紹

視頻生成是指利用人工智能技術，根據輸入的數據（如文本、圖像、視頻片段等）合成對應的視頻內容的過程。它可以從大量數據中學習視頻的規律和特徵，並基於這些規律和特徵生成新的視頻內容。

視頻生成是圖像生成的延伸和拓展。視頻生成不僅要生成每一幀高質量的圖像，還需要利用深度學習模型來理解和合成時間序列中的圖像幀以及相應的背景，從而生成連貫的符合物理世界規律的視頻流。

1.4 生成式人工智能的主要服務領域

生成式人工智能服務可以接受文本、圖像、音頻、視頻和代碼等輸入並生成多種形式的新內容，目前已廣泛服務於多個行業，包括以下幾例典型服務場景：

- **知識問答**：生成式人工智能根據其訓練數據中的知識和模式，結合外部的知識庫，能實時回應服務使用者的知識諮詢，提供準確、詳細的資訊解答，廣泛應用於智慧客服、企業內部知識庫、學術研究與教育等場景。
- **角色扮演**：生成式人工智能能通過對話系統模擬特定角色與服務使用者互動，如歷史人物、虛構角色、客服人員等，為服務使用者提供虛擬實境的體驗，可用於娛樂、教育、培訓等領域。
- **輔助寫作**：生成式人工智能能幫助服務使用者進行各種類型的寫作，如文章創作、故事編寫、論文寫作、文案策劃等，能提供創意靈感、語法糾錯、內容潤色、文本翻譯等功能，提高寫作效率和品質，廣泛應用於日常辦公、媒體、廣告、學術等領域。
- **輔助編程**：生成式人工智能可協助程序員編寫代碼、提供代碼建議、進行代碼審查、解釋代碼功能等，幫助提高編程效率，可應用於軟體開發、數據分析等領域。
- **數學推理**：生成式人工智能能進行複雜的數學計算和邏輯推理，可應用於解決數學問題、進行數學証明、提供數學解題思路和方法、輔助數學研究等。
- **人工智能代理**：生成式人工智能具有判斷決策能力，能感知環境、做出決策、並

採取行動，同時與其他系統進行互動。通常可以部署在各種複雜的應用環境中，廣泛應用於工業、醫療、金融等行業，如自動駕駛汽車、智能機器人、智能投資等。

- **人工智能藝術**：生成式人工智能在藝術創作領域得到廣泛應用，它能根據使用者指令或參考素材，生成風格各異的作品，用於藝術、設計等領域，提供創作靈感或輔助；能理解音樂理論和旋律結構，生成不同風格、類型的音樂，降低音樂創作門檻，助力音樂創作與配樂；還能生成動畫、短片等多種視頻，應用於影視、遊戲開發等領域，降低製作成本與門檻。

圖 1：生成式人工智能的主要服務領域

2. 生成式人工智能治理

2.1 生成式人工智能技術的局限和服務風險

2.1.1 技術局限

生成式人工智能模型在準確理解複雜語義，生成高質量內容方面成效顯著，能夠達到一般生活和生產需求水準。然而，生成式人工智能模型仍然存在諸多技術局限，稱為模型內生問題。這些模型內生問題最終導致模型生成有害內容。因此，指引鼓勵技術開發者、服務提供者和服務使用者了解以下模型技術局限以及其風險：

- **模型幻覺**是指模型生成的信息，與現實世界的實際情形不符，或與服務使用者的意圖相悖。問題的根源在於模型自身的生成機制。模型依據所學習的數據的分佈及模式通過算法的統計學習規則進行內容生成。在過程中，數據的局限性、算法的複雜性以及模型對語義理解的有限性，是模型幻覺產生的主要原因。按照當前技術發展水準，雖能採取多種手段對模型幻覺進行抑制，但尚未能徹底消除幻覺。以大型文本生成模型為例，模型幻覺會導致模型在回答問題或生成內容時，出現虛構、拼湊及移植等情況，從而產生與實際情況不符的回應。這不僅會嚴重影響模型輸出內容的可靠性和可用性，還可能帶來誤導決策或應用的潛在風險，因為這些決策或應用是基於模型的結果作出的。

- **模型偏見**是指模型在生成內容或作出決策時，存在的特定偏好及傾向。該倫理問題造成內容公平和公正性的缺失。模型偏見貫穿模型全生命週期：在訓練開發階段，算法設計缺陷導致算法偏見，訓練數據不全面、不均衡產生數據偏見，評估指標和方法不合理引發評估偏見。在實際使用階段，模型受過往數據影響而產生歷史偏見，受文化差異影響而導致文化偏見，對不同群體預判不公形成群體偏見，對個體特性敏感度不足引發個體偏見，受交互因素影響而導致交互偏見。此外，模型性能和數據時效隨時間變化產生時間偏見。

- **黑盒問題**是指生成式人工智能模型內部工作機制複雜且不透明，導致技術開發者、服務使用者和受眾難以理解模型如何生成輸出結果。這種生成機制的不透明性帶來了技術、倫理和實際應用上的諸多挑戰。例如，調試困難、優化受限等技術挑戰；責任歸屬、公平性和偏見等倫理與法律問題；透明度缺失、決策支持受限等使用者信任問題；模型對抗攻擊、濫用等安全風險。

- **數理能力**是生成式人工智能在數理邏輯推理方面顯示出技術潛力，但目前學界對於生成式人工智能是否只能模仿，或已具備一定科學性邏輯推理能力尚存在爭議。從效果上來看，生成式人工智能在需要運用數理邏輯解決的問題上表現

水準相對較低，即使是“點算”這類簡單的任務仍有可能出現錯誤。科學領域不斷嘗試探索生成式人工智能的數理能力。

- **對輸入變化的敏感性**指的是生成式人工智能模型對細微輸入變化反應過度的情況。即使是提示詞或查詢中的輕微差異，也可能導致截然不同的輸出結果，從而削弱模型的一致性與可靠性。
- **數據完整性**是建立可信生成式人工智能系統的基礎。具體風險包括：資料投毒（故意注入有問題或錯誤的數據）、資料漂移（模型所接觸的資料逐漸偏離訓練時的分佈），以及因保安控制不足而導致的未經授權修改。

為了降低這些風險，機構應當實施全面的數據管治措施，包括驗證技術、版本控制以及定期審計。妥善的安全控制和監察系統對於在整個人工智能生命週期中維護數據完整性至關重要。機構亦應就數據來源及處理方式提供適當的透明度，同時確保遵守相關法規。對於處理敏感資料的行業，建議考慮額外的保護措施，例如加強審計能力，以及在適當情況下採用能夠提供不可更改記錄的技術。

2.1.2 服務風險

建議中的人工智能管治框架建立了一個四層風險分類系統，根據潛在危害程度制定相應的監管措施，實現比例原則的監管方式。在最高層級，「不可接受風險」的系統，例如危害和影響市民安全的用途或潛意識操控等對社會構成生存性威脅的應用，應被禁止，而開發者亦須承擔法律責任。屬於「高風險」的應用，例如用於醫療診斷或自動駕駛等關鍵基礎設施的系統，必須通過合規評估，設有人為監督，並進行持續監察。至於對社會影響屬「有限風險」的系統，例如招聘工具或教育類人工智能，則需要履行透明度要求，提供用戶選擇退出的機制，並接受年度合規審計。最後，像垃圾郵件過濾器或創意工具等「低風險」應用，只需進行自我認證，監管負擔相對較低。這種層級式監管框架有助於在促進創新與加強保障之間取得平衡，確保監管力度與潛在風險的嚴重程度相匹配，涵蓋整個人工智能生態系統。

風險層級	定義	監管策略
不可接受風險	對社會構成生存性威脅的系統（例如危害和影響市民安全的用途、潛意識操控）	-全面禁止 -涉及開發或部署的行為須承擔法律責任
高風險	關鍵性基礎設施系統（例如醫療診斷系統、自動駕駛技術）	-必須進行合規性評估 -必須設有人類參與的監督機制 -實時監測和持續監控

有限風險	帶來中等社會影響的系統（例如人才招聘系統、教育人工智能應用）	-履行資訊透明的責任 -提供用戶選擇退出的機制 -每年定期進行合規審計
低風險	屬於最低風險類別的應用（例如垃圾郵件過濾系統、創意設計工具）	-企業自我認證

表格 1：風險分類系統

此外，生成式人工智能模型算法需要封裝成服務或產品的形式才能推向市場。例如 OpenAI 的 ChatGPT 實質是一種基於大語言模型的聊天服務，提供面向服務使用者的前端交互介面，後端可以選擇使用 GPT-4o、o1、o1-mimi 等多種不同的模型。生成式人工智能在服務階段仍然可能產生新的安全風險問題，指引稱之為服務衍生問題，如：

- **內容安全**是生成式人工智能服務面臨的關鍵問題。此類服務存在引發服務使用者創作、傳播不良內容，以及使受眾接觸不良內容的風險。例如，服務使用者可能借助生成式人工智能服務製造色情、暴力、血腥、恐怖、兒童虐待等危險內容；在互動時，生成式人工智能服務也可能向使用者灌輸不良價值觀，傳播仇恨、歧視性與煽動性言論等有害內容，讓使用者被動成為有害內容的受眾。這些不良內容經創作、二次加工並廣泛傳播後，會對受眾，特別是缺乏辨別能力的青少年產生隱性的消極影響，誘導他們做出違法犯罪或自我傷害等危險行為。

- **製造謠言**是指生成式人工智能服務能夠生成以假亂真的文本、圖像、音頻、視頻等多媒體內容。其低成本、易上手和快速使用的特性，可能被服務使用者有意圖、有選擇地製作成謠言，並大批量傳播，以達到混淆視聽、迷惑公眾的效果。一般受眾難以鑒別謠言，從而影響個人決策。隨著技術不斷發展，生成式人工智能被用於製造的謠言將更加逼真，為社會資訊環境的完整性帶來嚴峻挑戰。

- **模型越獄**指刻意繞過開發者設置的安全防護機制，使生成式 AI 服務可能被用於危險用途或濫用的行為。為防範生成式人工智能服務被用於危險用途和濫用，技術開發者為其設置了安全圍欄機制。正常狀態下，模型會識別並拒絕回應安全圍欄外的不安全請求。然而，在服務使用者端，針對安全圍欄的攻擊手段不斷湧現，此類攻擊手段被稱為模型越獄。例如，服務使用者可先輸入精心設計的攻擊指令，再提出非法請求，模型受攻擊指令干擾，原本應拒絕回應的非法請求，卻可能被正常處理，從而導致嚴重的安全隱患。

- **數據洩露**指生成式人工智能服務（尤其是聊天對話服務）可能會以各種形式收集服務使用者的資料，包括服務使用者主動提供的資料、上傳的文檔以及通過設備讀取的個人資料，在數據的傳輸和處理過程中可能導致個人或企業服務使用者的私有數據對外洩露。

在防範潛在風險與保障言論自由之間取得平衡，始終是執法部門面對的一項持續挑戰。雖然當局致力於維護受基本法保障的言論權利，但當人工智能技術被濫用於非法活動時，明確界定法律邊界就變得尤為必要。此類被禁止的應用包括：散播助長暴力行為的指引（例如製造爆炸裝置或施襲手法）、生成猥褻內容、製作經過合成操控的媒體內容或虛假資訊以協助詐騙活動或擾亂公眾秩序，以及開發旨在入侵資訊系統的惡意程式碼。

2.1.3 模型生命週期和人為監督

一個生成式人工智能系統的生命週期可分為以下幾個階段：

（1）規劃與數據獲取；（2）模型開發；（3）部署與集成；及（4）使用與維護。

針對生成式人工智能模型所固有的內生風險，開發者、機構與個人在追求實用性的同時，應充分考慮風險代價，從而以負責任的態度在各階段開發和利用生成式人工智能。下文將說明生成式人工智能模型生命週期中存在的風險，並提示相關方應盡的責任。

2.1.3.1 規劃與數據獲取

生成式人工智能旨在創建新的內容，比如文本、圖像、音訊、視頻，這需要它能夠類比人類創造力，並在一定程度上超越現有的資料收集和使用模式。生成式人工智能常採用自監督/無監督學習，通常需要比傳統人工智能更大的資料量，因此面臨更高的資料風險，例如：

- **有害資料風險/資料投毒**：在採集流程中，因收集到錯誤的資料、異常值或攻擊者特殊設計的樣本，模型學習到錯誤的資料分佈模式或決策邊界，從而導致模型性能下降、產生不適宜的結果或偏頗的模型行為。
- **數據偏見風險**：由於資料來源、採集方法及採集時間的差異，採集資料的客觀性與全面性缺失，模型表現出一種偏見特性。這些模型會常常重複並放大存在於其訓練資料中的偏見。
- **個人資料私隱風險**：技術開發者在收集資料時，可能無意或有意地接觸敏感性個人資料，從而侵犯個體的私隱。
- **知識產權風險**：訓練生成式人工智能模型及使用生成內容可能涉及知識產權風險，無意中使用受版權保護的材料可能會導致侵權索賠，而使用與現有受保護作品極相似的生成內容可能導致增加法律風險。

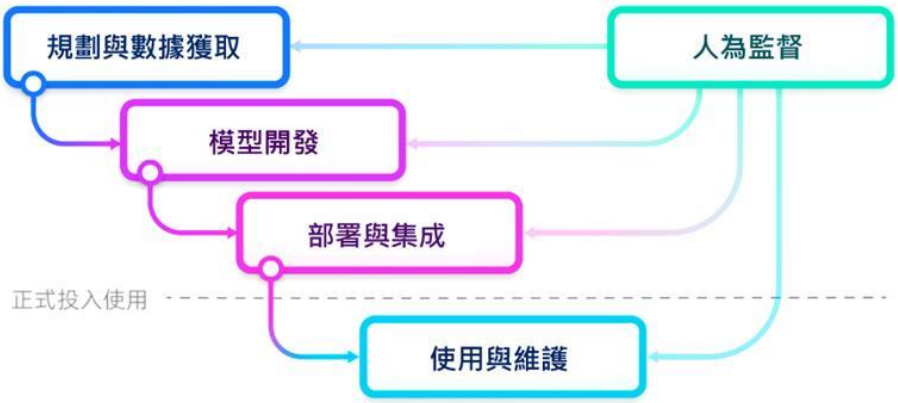

圖 2：生成式人工智能系統的生命周期

為了降低這些風險，在規劃及收集數據階段，評估數據來源及其質量至為關鍵。應採取措施確保數據收集與清洗的標準化。具體而言，數據來源應遵循多元化原則；透過自動化或人工方法分析數據分佈，並過濾有害內容；同時，在收集數據時必須獲得相關持份者的適當授權，以避免侵犯個人私隱。

針對法律、監管、政府及銀行等高風險行業，進行特定領域的精細化調優尤為重要。由於這些領域往往涉及語言表達細緻、法律精準度要求高以及資料保密等因素，通用模型未必能有效滿足需要。通過針對性調優或使用精選的領域專屬數據集（如立法檔案或監管框架）進行訓練，可有效提升模型的準確性、合規性及可信度。

具備多模態能力的生成式人工智能，能夠大幅提升公共服務的提供效率。例如，將生成式人工智能語音助手整合至政府部門的通訊渠道中，可協助應對人手短缺問題，有效處理日常查詢，同時讓人類職員專注於複雜或敏感個案。憑藉語音識別、情感分析及語音轉文字摘要等功能，這些人工智能系統能夠將大量互動數據轉化為可操作的洞察，既有助於即時回應公眾需要，亦可支援長遠的政策制定。

2.1.3.2 模型開發

模型開發包括模型體系結構的設計和選擇，訓練算法的規劃，設計和實現及評估等過程。首先，我們要評估生成式人工智能模型是否契合業務目標至關重要，要有流程闡明決策依據。在規劃階段，想要確保生成式人工智能項目的成功，關鍵在於了解不同模型所需資料集，選對學習算法，同時釐清其可擴展性、存儲及開發時間方面的限制。模型開發流程是非線性過程。隨著新資料的到來或業務需求的變化，模型可能需要再次訓練或更新。在生成式人工智能模型開發過程中，可能面臨如下風險問題：

- **技術風險**：根據不同的模型架構和能力，模型可能過於緊密地匹配訓練資料（過擬合），使得模型的生成內容與訓練內容過於相似，從而失去了其普遍性；或者模型未能充分學習到訓練資料中的模式（欠擬合），導致模型無法輸出滿足指令要求的內容。這兩項技術風險都會影響模型的整體準確性和可靠性。

- **知識產權風險**：模型開發可能涉及採用第三方架構或預訓練權重。如在過程中侵犯他人的知識產權 (如專利、版權) 或違反相關授權要求，將可能引致知識產權爭議。

所以，在模型開發階段，模型開發者應充分了解所選用模型的準確性與可靠性，應設計適當的指標，對模型的技術風險進行充分評估。若採用開源模型架構或權重，還需要遵循開源許可證要求。

2.1.3.3 部署與集成

部署與集成是服務使用者使用人工智能模型前的最後一個階段，服務提供者需要準備足夠的計算資源來支援模型的運行，並集成直觀易用的服務使用者介面，讓服務使用者可以輕鬆地與模型進行互動。此階段應重點關注安全風險：

- **系統集成風險**：在多個組件整合為一個工作整體的過程中，可能會因相容性情況、組件故障及安全性漏洞等導致整個系統處於不穩定狀態。
- **許可權控制風險**：由不同開發者或第三方開發的組件，若存在不當的權限繼承，可能導致權限無意間擴散，進一步引致知識產權侵權或資料違規使用等問題。

為了應對這些風險，服務提供者應建立具備冗餘及高可用性的系統，確保即使單一組件發生故障，亦不會影響整體安全性；設計全面的系統整合測試策略；以及明確界定各系統組件的訪問權限，並實施嚴格的存取控制政策，以限制對關鍵組件和服務的訪問。

機構亦應評估部署策略，在採用生成式人工智能的同時兼顧安全考慮。由於人工智能提示詞中可能包含敏感的商業數據或個人資料，機構在選擇服務模式時(無論是基於雲端的解決方案、自建的開源模型，還是裝置端的部署方式) 均應仔細評估相關風險。對於許多情況而言，採用混合部署模式，按應用場景的重要性及資料保護需求作出部署選擇，或許是較為合適的方案。

採用互操作性標準，例如開放式 API 及安全數據共享協議，能促進政府平台與私營機構的人工智能解決方案之間更順暢的整合。透過促進有結構的協作及技術兼容性，這種做法有助加快人工智能技術的應用進程，減少重複投入，並促進開放創新的生態系統。

與國際框架保持一致的監管安排，有助提升信任並促進跨境人工智能合作。與其他司法管轄區已建立的廣泛認可標準保持一致，能為全球合作夥伴及投資者提供更高的可預測性和信心。清晰劃分與人工智能相關問題關聯的法律及監管機制，包括由哪些主管機關負責處理各類人工智能事務，亦有助提升制度透明度及管治效能。

2.1.3.4 使用與維護

當生成式人工智能服務正式投入使用，服務使用者通過互聯網或在區域網絡向模型

提供輸入以獲得期望內容輸出。在向公眾領域提供服務時，需要特別關注以下風險：

- **內容安全風險**：根據服務使用者設計的不當輸入，生成式人工智能可能生成不符合法律和道德要求的不良內容和謠言。
- **資料洩露風險**：根據服務使用者提供的提示生成回答時，生成式人工智能有可能無意間包含服務使用者的個人資料或其他機密資料，或存取檢索增強生成（RAG）過程中使用的資料。
- **版權風險**：由生成式人工智能服務所產生的內容，其版權歸屬（如適用）並不總是明確無誤。不同司法管轄區對於此類內容的版權存續及歸屬問題，法律規定可能有所不同。因此，必須謹慎處理相關版權問題，包括明確界定生成式人工智能服務輸出的內容是否享有版權及其歸屬權人，以及防止相關輸出內容構成版權侵權行為。
- **可信性、倫理和社會風險**：生成式人工智能能夠創造逼真但虛假的內容，如假新聞、偽造的音訊視頻或與倫理及社會價值相悖的內容，對社會信任構成威脅。

技術開發者有責任採取技術手段對服務使用者輸入內容進行過濾，避免因有害意圖而生成不良內容，並採取相關措施對生成內容進行標識，從而和真實內容作區分。服務提供者應當提供清晰的操作指南和技術文檔給服務使用者，說明如何正確使用；在可能的範圍內，應向服務使用者聲明生成內容的版權歸屬；只有在徵求服務使用者同意的情況下，才可對其使用活動作記錄（日誌），並必須在不違反個人資料的情況下進行存儲。服務使用者應當充分意識到人工智能生成的內容可能存在虛假性，並要主動查核內容的真實性。

2.1.3.5 人為監督

適當的人為監督對於確保生成式人工智能系統的信賴責任框架至關重要。人為監督生成式人工智能的程度，要依據對其各個環節（如數據收集、模型訓練、生成結果）的影響評估來決定。如果某個環節對生成式人工智能的影響越大，就越需要加強人為監督。根據人為監督程度，可將模型分為兩類：

- **生成式人工智能協作模型**：對於那些僅用於影響較小決策的生成式人工智能協作模型，由於人類的判斷能夠起到有益的補充作用，且資料量有限，所以需要一定程度的人為監督。
- **人工主導模型**：當生成式人工智能協作模型不足/適用時，就應該採用人工主導的模型，也就是以人工決策和操作作為主要的運行方式，人工智能起輔助作用。

2.2 治理的五大維度

為促進生成式人工智能的有效且有益使用，指引從五個維度介紹生成式人工智能治理框架。在此框架下，持份者應明確行為範圍，精準評估潛在風險。這五大維度為：

- **個人資料私隱**
- **知識產權**
- **犯罪防治**
- **真實可信**
- **系統安全**

2.2.1 個人資料私隱

生成式人工智能的快速發展，為個人資料私隱保障帶來前所未有的挑戰。由於人工智能系統本身的高度複雜性，在開發的各個階段均可能對個人資料私隱及敏感資訊的安全構成風險。從數據收集、處理，到模型訓練、優化，再到最終應用及服務提供，即使是微小的疏忽，也有可能令敏感資訊暴露於潛在威脅之中。

為了應對人工智能訓練階段的個人資料私隱問題，聯邦學習（Federated Learning）等創新技術應運而生。這項技術容許模型在多個去中心化設備或伺服器上進行訓練，這些設備或伺服器持有本地數據樣本，而無需進行數據交換。系統僅共享模型更新，有助於在有效推動模型開發的同時，盡量降低私隱洩露的風險。

關於個人資料私隱，個人資料收集的目的及方式、個人資料保存的準確性及期限、個人資料的使用、個人資料的安全性、個人資料政策及做法的公開及透明度，以及個人資料的查閱及更正，是生成式人工智能開發及服務提供生命週期中的關鍵部分。

因此，確保在生成式人工智能開發及服務提供的整個生命週期內，個人私隱及敏感資訊的安全性，至關重要。這不僅關乎保障個人權利，也是維持公眾對生成式人工智能技術信任，以及推動行業健康可持續發展的關鍵。

2.2.2 知識產權

各地的知識產權制度為創作者和發明人提供一套專門而平衡的法律框架，以保障其合法權益，並促進創意和創新的文化。然而，生成式人工智能的迅猛發展，正從數據收集、模型訓練到內容生成等多個層面，為現有的知識產權制度帶來前所未有的機遇與挑戰。

在數據收集及模型訓練階段，使用受版權保護的材料進行人工智能訓練，引起了社會各界對潛在侵權行為及版權豁免適用範圍的廣泛關注。這促使學術界、業界及法律界積

極展開深入討論，並探索是否需要制定明確條文，為模型訓練提供專屬的版權豁免規定，藉此在促進創新與保障創作者利益之間尋求平衡。

2.2.3 犯罪防治

生成式人工智能的發展，為預防及打擊犯罪帶來機遇與挑戰。一方面，人工智能的應用能夠大幅提升執法效能，令傳統的執法方式轉型為更精密、數據驅動的模式。然而，相關治理工作必須超越單純的法律框架，採取更全面的方式，納入倫理考量、社會影響、公眾接受程度及社區反應等因素。實施相關技術時，須保持對其能力與局限性的透明度，以維護公眾信任，並確保技術的應用符合社會價值觀及大眾期望。

同時，生成式人工智能亦帶來嚴峻的治理挑戰，因不法分子正在積極利用相關技術。由生成式人工智能驅動的深度偽造技術，能夠製作極為逼真的虛假音訊及視訊內容，模仿個人的外貌和聲音。這些具高度欺騙性的內容正日益被用於散播虛假資訊、操控公眾輿論，以及在通訊或電子交易中冒充他人以實施詐騙，對公共安全、私隱保障及公眾對數碼資訊的信任構成多重威脅。

2.2.4 真實可信

生成式人工智能的可信度，是指其具備穩定可靠的性能表現，能夠在各類應用場景中持續且精準地輸出符合預期、真實可靠的結果。對生成式人工智能系統而言，真實可信是生成式人工智能信賴責任的核心要點之一。這一責任涉及的任務是，構建一套科學、嚴謹且行之有效的機制與框架，來確保生成式人工智能開發者、營運者以及使用者對模型運行邏輯、行為模式及其所產生的廣泛影響承擔相應責任。特別是在系統生成資訊出現失實、偏差或誤導性內容時，能夠依據該機制和框架，清晰、準確地劃分各相關主體的責任。

當前，人工智能生成內容的信賴責任正面臨著前所未有的重大挑戰。生成式人工智能所蘊含的技術架構具有高度複雜性，其內部運行邏輯與算法機理存在一定程度的不透明性，這使得在實際應用過程中，對其決策過程的追溯與解析變得極為困難。目前已知能夠對生成式人工智能的真實可信產生影響的問題包括，算法設計環節存在缺陷、訓練數據存在偏差或雜訊，以及運行過程中遭遇突發異常狀況等。由於難以精確判斷錯誤資訊產生的根源，就無法在發現問題時迅速、準確地確定責任主體，這嚴重影響了生成式人工智能的可信賴程度與應用推廣。

2.2.5 系統安全

生成式人工智能的治理框架下，系統安全是核心要素，其重點在於保障系統本身以及資料不被未經授權者訪問和破壞。然而，敏感資訊處理過程中出現的特殊漏洞和模型面臨的逆向攻擊風險正在不斷增加生成式人工智能的安全隱患。同時，資料投毒攻擊也是一個突出問題。惡意攻擊者會蓄意篡改訓練資料，干擾模型的學習過程。一旦資料投毒成功，

模型就可能做出錯誤或有偏差的決策，這不僅會破壞人工智能應用的完整性，還會極大削弱用戶信任度。

為應對這些挑戰，技術開發者和服務使用者需要建立精細的監控和更新機制，嚴格執行安全措施，以此確保資料的完整性和可信度。在防範資料投毒方面，應實施嚴格的資料驗證流程，利用異常檢測算法識別和篩選可疑資料條目，同時保證訓練資料集來源可靠、品質優良。此外，定期對資料來源進行審計，並構建安全的資料傳輸通道，能夠進一步降低資料投毒的風險，為生成式人工智能的安全穩定運行提供堅實保障。

2.3 治理的主要原則

2.3.1 遵守法例

在生成式人工智能技術從研發、服務供給到實際應用的全流程中，所有持份者均需以高度的法治意識和嚴謹態度，嚴格遵循法律條例。就香港而言，生成式人工智能技術運作的每一個環節，毫無例外地都必須與香港現行法律法規的具體規定和精神實質完全契合，不容許有任何偏離或違規操作。具體而言，技術開發者在模型訓練語料獲取過程中應保護知識產權和個人資料，而模型不應生成含有法律法規禁止，違背公序良德以及損害知識產權和個人資料的內容。若生成式人工智能服務覆蓋的行業、地區和國家有更高要求，技術開發者、服務提供者和使用者都應予以尊重並遵守。服務提供者和使用者應理解自己的社會責任與法律義務，防止如傳播虛假或有害內容。

2.3.2 安全透明

在生成式人工智能的發展過程中，必須同時針對源自模型本身及服務層面衍生的問題進行應對與解決，以降低技術的不安全性及不透明性。在模型層面，應透過算法優化及數據治理，消除涉及違法、違規或違背倫理標準的有害內容。在服務層面，服務提供者必須充分向用戶披露相關風險，並運用如加密技術及可解釋人工智能等手段，提升模型的安全性與透明度，確保技術的可靠運行。

一些市場上的開源模型為例，其推理技術為內容生成賦予清晰的邏輯脈絡，能夠生動地展示內容生成過程中的關鍵步驟與邏輯。這在一定程度上打破了生成式內容「黑箱」式運作的局限，使用戶能夠直觀理解內容生成的依據，大大提升了生成式人工智能的透明度，為技術的安全可靠應用奠定了堅實基礎。

開源模型通常具有更高的透明度，公眾可審查其訓練方法、數據來源及算法設計，從而有助於更廣泛地驗證安全措施及發現潛在偏見或漏洞。而專有模型雖然往往擁有先進的功能，但由於商業考慮，其透明度存在固有限制，使得外界獨立驗證變得更具挑戰性。

此類比較可為持份者提供有關不同開發模式下透明度取捨的重要洞察，並為完善治理框架提供有力參考。

圖 3：生成式人工智能治理的主要原則

2.3.3 準確可靠

在生成式人工智能的研發與應用進程中，需在模型開發與服務階段謹慎管理，以妥善應對各類風險。在模型開發階段，技術開發者應充分利用先進技術與科學方法，著力降低模型幻覺等內生問題，從而減低對後續應用的潛在影響。例如引入 RAG 技術，通過在生成內容時從海量外部知識源精準檢索關聯資訊，並將其有機融入生成過程，以此增強模型對真實世界知識的理解與運用，確保生成內容的準確性與可靠性，避免因模型幻覺導致生成內容偏離事實。同時，依據服務的具體形式，精心設計並提供操作簡便、高效實用的事實查證途徑，如搭建權威資料檢索介面、開發智能比對工具等，助力使用者進行人工查證，切實提升生成結果的可靠性，全方位保障生成式人工智能服務的安全性、穩定性與合規性，推動生成式人工智能技術在可控、可信的軌道上穩健發展。

2.3.4 公平客觀

生成式人工智能服務需深度貫徹多元化與普世性原則，全方位規避資訊傳播失衡，堅決杜絕因某類資訊過度集中而催生的資訊回音室問題。從語料數據獲取的源頭，到內容生成的各個關鍵環節，都要嚴格把控，以消除模型偏見。在語料篩選時，要廣泛涵蓋各類資訊，確保不同領域、不同背景的資料均有充分體現。在模型訓練和內容生成過程中，通過科學算法和嚴格審核機制，避免生成內容出現針對信仰、政見、國別、地域、性別、年齡、民族、膚色、行業、健康狀況、收入水平以及生活習慣等方面的偏見與歧視性表達，從而為用戶提供公平、客觀、包容的內容，推動生成式人工智能技術在促進資訊公平與社會和諧方面發揮積極作用。

2.3.5 實用高效

生成式人工智能作為創新且極具生產創造力的技術，正深刻變革各領域運作模式。技術開發者與服務提供者肩負著持續提升其效用的重任，需確保生成式人工智能能夠精準、高效且優質地生成契合服務使用者意圖的內容。一方面，通過優化算法和模型架構，提高內容生成的準確性與相關性，杜絕無效或偏差資訊；另一方面，應透過探索該技術在不同任務、複雜場景及多元產業的應用，充分發揮其潛力。生成式人工智能透過解決現實問題並大幅提升工作效率，將能推動產業升級，促進社會進步和民生改善。憑藉創新應用，生成式人工智能可成為產業與社會發展的驅動力。

3. 生成式人工智能技術開發者、服務提供者和服務使用者實用指南

綜合上述討論，我們就生成式人工智能的技術開發者、服務提供者及服務使用者三方，根據其於持份者責任矩陣中所界定的角色、權利、責任及義務，提出以下建議：

持份者	定義	責任
技術開發者	從事生成式人工智能系統基礎模型及算法的開發、訓練及維護的機構及個人。	-道德規範的模型開發 -技術保障措施的落實 -持續監督與監察 -數據權利管理
服務提供者	作為開發者與用戶之間的中介，負責將生成式人工智能技術部署為面向客戶的應用或服務的機構。	-內容治理 -私隱保護 -問責措施 -用戶數據處理
服務使用者	出於個人或專業用途而使用生成式人工智能服務的個人或機構。	-道德使用 -意識提升與自我控制 -社群保護 -內容核實

表格 2：角色定義矩陣

3.1 技術開發者：建立完整的開發團隊和正確的工作方式

成立具備良好架構的生成式人工智能開發團隊，並採納健全的工作實踐，對於推進技術創新與確保安全合規同樣重要。開發者應恪守重視知識產權保障的原則，並兼顧數據完整性與中立性等其他關鍵因素。

- **設立數據團隊及負責人**：專責的數據團隊應由指定負責人領導，確保遵守相關法律及其他相關法規，避免任何侵權或違規行為。除了負責管理數據分佈及保持公正，防止歧視外，該團隊亦須在模型微調過程中，確保數據集符合正確的價值觀及事實準確性，並建立健全的質量控制機制，以維持上述標準。此外，數據的獲取、存儲、處理及傳輸，必須遵從相關法律法規，包括個人資料私隱及網絡安全方面的要求。必須採取嚴格的安全措施，以防範數據洩露及濫用。透過落實以上措施，數據團隊將能有效支持負責任且符合道德標準的人工智能發展。
- **設立算法工程團隊**：在算法設計和優化中提高對安全性和可信性的重視，確保

技術在預訓練、微調和推理階段的安全性，及時發現和修復安全性漏洞，並儘量採用推理技術來提高生成內容的可解釋性。同時，要在系統層面上，加強對生成內容可靠性的技術保證。採用如 RAG 和知識庫等技術來保證生成內容的時效性和準確性。

- **設立質控團隊**：全面測試是確保人工智能應用程式安全部署且功能正常的關鍵手段。建立正式的測試要求以驗證模型抵禦安全威脅的能力—包括對抗性攻擊能力和對資料洩露的敏感性—尤其重要。讓服務使用者參與測試過程，從服務使用者的角度識別安全問題。而標準化的測試基準有助於在候選模型之間進行有效比較。為此，系統必須有具體的性能指標。開發者應根據性能指標 (例如精確度、召回率和準確性) 制定清晰、可衡量的實現路徑，並制定可以對模型的運行進行人為監督及介入的控制措施或手段。此外，必須實施基於異常的報告系統，以突出顯示性能偏差，並定期進行評估以保持模型的準確性和有效性。最後，技術開發者必須定期審查測試結果，進一步保障測試結果的準確性，以及保證模型和產品的穩定性和可靠性，確保在各種應用場景中能夠正常運行和輸出可靠結果。

- **建立營運原則**：透明度與可解釋性至關重要，開發者應在可行情況下披露訓練數據來源、模型架構及評估指標。機構應制定相關政策，規定何時可接納人工智能生成的內容，例如要求用戶在使用前仔細核對人工智能生成的資料、驗證參考來源及確保內容準確。必須優先考慮數據質量，選用高質素、多樣化、具代表性並定期更新的訓練數據。人工智能系統應為輸出提供顯明的解釋，並內置來源核查、事實查證及驗證機制。持續監察及定期審計至為關鍵，以便及早識別及糾正錯誤、偏見或不準確之處。應積極鼓勵用戶反饋意見，以提升系統可靠性，同時開發人員及內容創作者亦須具備相應領域專業知識。必須建立問責機制，以應對錯誤資訊的傳播，並制定行業層面的標準，以確保一致性及準確性。與本地及國際的學術機構及研究單位保持緊密合作，對於掌握最新技術進展及利用專業知識至為重要。最後，行業領袖與監管機構應共同制定涵蓋數據質量、透明度、問責性及偏見緩解的統一指引。

- **設立合規團隊**：對研發的模型和產品定期進行合規審查和評估，從而避免對社會秩序、公共安全和道德倫理造成負面影響。建立完善的文檔制度，促使技術開發者遵循透明度原則，公開技術原理和使用規則，從而使服務使用者和監管機構能夠理解和監督其技術應用，建立信任，減少技術被濫用的風險。

- **對技術開發者及服務提供者應遵從更高的標準**。對於高風險的人工智能生成內容，如深度偽造、身份證明文件圖像及財務材料，應加設不可移除的浮水印或嵌入式標識碼，以確保可追溯性及問責性。透明度至為關鍵—人工智能模型必

須以經過驗證且可靠的來源進行訓練，服務提供者亦應整合可信的參考資料，並公開披露其信息來源。當人工智能系統輸出的內容無法被驗證時，應主動向用戶發出提示，以防止未經證實或虛假內容的流傳。準確性的責任應由開發者及服務提供者承擔，而非轉嫁予用戶。

- **獨立評估機制**應適用於服務提供者，同時亦應自開發階段起涵蓋技術開發者。開發者應接受涵蓋內容準確性、偏見、有害輸出及私隱合規等風險範疇的審計。作為主要持份者，開發者必須承擔更大責任，以確保人工智能安全。落實上述措施，將有助於強化問責機制及透明度，並提升指引在防範人工智能濫用方面的整體效能。

3.2 服務提供者：建立負責任的服務框架和服務建設流程

3.2.1 建立負責任的生成式人工智能服務框架

生成式人工智能服務提供者要確定可以帶來顯著價值的具體業務或機會，並根據服務可行性、策略目標一致性和潛在影響來決定服務提供的優先順序並建立負責任的生成式人工智能服務框架。我們提出以下建立框架要考慮的方面：

- **保證服務合規**：服務提供者選擇和使用的基座模型應符合香港的相關法規和社會道德標準。服務提供者有責任保證其服務系統不輸出違法、違規或不當的內容，應當建立機制提升系統的可追溯性與可審核性，有效減小被輸入惡意資料的風險；對生成的圖像、視頻等內容進行標識；對不適合輸出或輸出存在偏見的內容，以及對主要服務對象為特殊群體如未成年人的服務，應加強風險通知。

- **保證資料安全**：服務提供者要遵守《個人資料（私隱）條例》(《私隱條例》)（第486章）等相關資料保護法規，在收集、處理、使用、存儲、保留和刪除個人資料等的服務使用者資料時，充分保護服務使用者私隱權益，避免過度收集、濫用或外洩服務使用者資料，同時加強對敏感性資料進行加密和脫敏處理，確保資料在流轉過程中的安全性和私隱性；必要時與技術開發者緊密合作，並在服務使用者中開展資料安全調研，及時發現和修復安全性漏洞。

  服務提供者必須採取具體措施，有效管理個人資料私隱。首先，應標準化跨行業的數據保護協議，以確保數據完整性，並促進國際數據流通與合作，服務提供者可參考不同指引，包括個人資料私隱專員公署（私隱專員公署）發出《保障個人資料：跨境資料轉移指引》。其次，必須落實強而有力的技術保障措施，例如先進的匿名化技術及強化加密技術（即從數據集中刪除或加密個人識別資訊），以保障數據安全，防止被重新識別。第三，務必遵守適用本地及國際的數據保

護法律 (如《私隱條例》及歐盟《通用數據保障條例》)，並取得服務使用者的明確同意，保障其合法權益。最後，服務提供者應持續進行監察與評估，以應對不斷演變的數據安全威脅，確保相關措施持續有效並與時俱進。

- **保證系統安全**：服務提供者需對系統的安全性進行持續監測和評估，制定系統和資料遭受攻擊的防範措施和應急處理方案；系統在功能或運行方式上有重大變化時，應重新評估以識別和防範新的風險。

- **保證系統的可信性**：服務提供者需要遵循透明度原則，公開其服務背後的技術原理，說明服務使用者和監管機構理解和監督其提供的服務。提供詳細的服務說明和使用指南，使得服務使用者能夠正確理解和使用服務。服務提供者要制定明確的信賴責任框架和定期的審查週期，與持份者一起進行監察，從而營造合規文化，提高服務的可靠性。此外，聘請獨立審計師，定期對服務的品質，安全性和合規進行審計；聘請倫理政策專家，以增強服務水準與道德標準和政策的一致性。

3.2.2 負責任的服務建設流程

- **服務採購**：服務使用者在採購生成式人工智能服務時，經濟和安全因素往往是重點考量對象。一方面，價格需在預算範圍內，成本效益也要合理；另一方面，資料安全、私隱保護以及系統穩定性等安全指標，直接影響著服務的質量和使用體驗。因此，服務提供者應建立明確的財務協議與服務安全協議，以保障服務的有序開展。

  另外，要建立獨立評估機制以及妥善的檔案記錄。獨立評估能從客觀中立的角度審視服務，發現潛在問題；而完善的檔案記錄則詳細記錄了服務的各個環節，不僅增強了透明度，也為信賴責任框架的構建提供了有力支撐。

- **風險評估**：生成式人工智能服務提供者要在服務建設的各個階段從多方面入手開展服務風險評估。評估方面包括：在資料層面，審查訓練資料品質，排查缺失、錯誤標注與重複等問題，分析資料偏差，確保資料安全，防範私隱洩露、篡改與濫用。在算法和模型層面，檢查算法是否存在邏輯與安全性漏洞，評估模型決策過程的可解釋性以及在不同輸入條件下的穩定性。在應用場景層面，確保服務符合各場景的法律法規與行業標準，分析服務的缺陷，如虛假資訊，對使用者體驗的負面影響，考量其社會風險。在營運和管理層面，評估服務基礎設施與運維能力，關注人員操作與道德風險，評估協力廠商供應商的可靠性與穩定性。同時，要確保評估的獨立性，並建立妥善的檔記錄機制。獨立評估能從客觀中立的角度審視服務，發現潛在問題；而完善的檔記錄則確保服務全流程的透明度，為信賴責任框架的構建提供有力支撐。

- **試點項目**：生成式人工智能服務提供者在推廣大規模服務之前，應開展小規模的試點專案，在明確目標與範圍的情況下，驗證其在特定業務場景的可行性，考量對業務效率和成本的影響，界定業務流程、使用者群體和資料邊界。這樣的試點根據目標和資料特點選擇合適模型，評估性能、可擴展性和成本，必要時對模型進行微調。

- **服務維護**：服務維護包含了對服務的品質，安全性和用戶滿意度的可持續性的改善。服務提供者必須優先考慮與技術開發者和服務使用者的透明溝通，以建立信任並確保合規。這包括披露生成式人工智能服務的使用規則、目的、益處和局限性，同時尊重資料主體和使用者的權利，並告知個人其個人資料的使用方式。服務提供者應努力提高生成式人工智能服務透明度，提供對決策過程的洞察，並建立回饋管道，以使用清晰易懂的語言促進包容和負責任的生成式人工智能服務過程，改善使用者體驗。

3.3 服務使用者：人工智能有益向善的主動維護者

生成式人工智能服務使用者在使用過程中承擔重要責任，需提高對技術與服務安全風險的認識，從而自覺地成為合規、安全的人工智能生態環境的建設者。我們對生成式人工智能服務使用者提出以下建議：

- **合法合規使用**：按照服務提供者和監管機構（例如私隱專員公署）的指導和要求合理使用生成式人工智能服務，不得將服務用於違法、違規或不當的目的。使用者通過生成式人工智能服務生成內容，須遵守香港的法律法規，包括但不限於與版權、私隱和反歧視的相關法律等;如果生成內容存在潛在或明顯違規、與事實不符等問題，服務使用者應及時向服務提供者和有關主管部門回饋報告，移除相關內容，避免傳播違法、違規或不當的內容是正確使用生成式人工智能的基本要義。

- **堅持自主判斷**：生成式人工智能是我們的工具和助手，而不是我們的替代者。生成的內容可以成為我們工作和生活的參考，但不應在不經我們驗證和判斷的情況下被採用。服務使用者應注意，人工智能生成的內容可能包含誤導性、虛假或不準確的資訊，作為負責的使用者，我們應當具備法律、道德與風險管理等多方面的意識與能力，能夠通過驗證和審查生成內容，作出獨立的資訊判斷。

- **了解責任和義務**：服務使用者在使用任何生成式人工智能服務之前，應仔細閱讀並熟悉相關平台或軟體的使用條款，了解自己的責任和義務。這些條款往往涵蓋私隱、安全、道德規範以及法律合規等主題。例如，明確規範服務使用者不得指使人工智能生成任何帶有仇恨、歧視、誹謗或其他不道德與違法內容，以

及對使用者的權益的保護，如對個人資料的共用範圍的限制，避免服務提供者超範圍記錄、傳播使用者個人資料。

- **引用說明**：為確保透明度、信賴責任框架和安全，服務使用者應明確說明是否使用了生成式人工智能參與內容生成或決策，並且必須承擔其道德和法律影響的責任。

- **保護私隱**：服務使用者應熟悉生成式人工智能服務的私隱政策，以了解其資料收集、使用和共用的具體做法。建議選擇不將共用資料用於生成式人工智能訓練的服務，並避免傳輸敏感的個人資訊，同時採用假名化或匿名化手段來保護私隱。定期審查和刪除生成式人工智能服務中的資料，如果在生成的內容中發現不準確之處，服務使用者應提交更正或刪除請求。

- **謹慎傳播**：任何由生成式人工智能系統所生成並進一步散播的內容，一旦對社會、經濟或文化產生影響，最終的責任仍落在內容傳播者身上。這意味著服務使用者需要對內容可能造成的誤導或負面後果加以評估並承擔責任。服務使用者需主動檢查生成內容的真實性、合法性和適宜度，並在必要時尋求專業建議或進行二次審查，以降低潛在風險。同時，若服務使用者有意發布人工智能生成的內容，則應在公開時披露其來源。尤其當此內容牽涉到商業用途或大規模傳播時，披露來源更為重要。這種資訊透明度不僅有助於增進公眾對生成式人工智能的信任，同時能夠確保自身內容創作和傳播的合法性。

- **尊重知識產權**:為了維護生成式人工智能生態系統並鼓勵服務使用者合法創作，必須自覺尊重知識產權。這包括採用必要的技術措施及解決方案，避免生成的內容構成對受版權保護作品的全部或實質性複製，以防止引發版權糾紛。具體做法可包括搜尋及評估生成內容是否構成侵犯版權、商標或專利權。如發現生成內容涉及侵犯他人知識產權，應根據適用的行政及 / 或法律措施（例如使用條款、下架通知及 / 或法院命令）及時刪除或修改有關內容。

致謝

機構：

香港科技大學

香港生成式人工智能研發中心

主筆：

韓斯睿、郭毅可、黃紅英

韓斯睿教授，香港科技大學助理教授及RGC-富布萊特研究學者，同時擔任香港生成式人工智能研發中心預研部主任。

郭毅可教授，香港科技大學首席副校長及講席教授，同時擔任香港生成式人工智能研發中心主任。

黃紅英博士，香港科技大學特別顧問，同時擔任香港生成式人工智能研發中心首席營運官。

附錄

隨著生成式人工智能技術的發展，各國和各行業都開始加強對其治理和監管，以確保技術的合法合規。在為技術開發者、服務提供者和服務使用者提供具體行動指南之餘，指引整理各國和各行業對生成式人工智能治理的相關要求，以為他們提供政策參考資訊，幫助他們在全球範圍內更安全和合規地開展生成式人工智能活動。此外，香港人工智能行業治理特色鮮明，指引總結香港可信生成式人工智能行業應用原則，來為相關方提供參考，參考內容可見附錄。

1　特定國家和地區的生成式人工智能治理要求

1.1　國內

國家在生成式人工智能領域的投入與發展成效顯著，已成為全球人工智能發展進程中的關鍵力量。國家將人工智能治理提升到關乎全人類命運的高度，認為這是世界各國共同面臨的重要課題。基於此，國家發布《全球人工智能治理倡議》，呼籲各國秉持共同、綜合、合作、可持續的安全觀，堅持發展與安全並重，通過對話與合作凝聚共識，構建開放、公正、有效的治理機制，讓人工智能技術為人類謀福祉，推動人類命運共同體的構建。

國家一直以來穩步推進人工智能治理與監管。此前已針對推薦算法、深度合成等技術服務制定了針對性管理要求。在生成式人工智能領域，國家互聯網信息辦公室聯合相關部門在《網絡安全法》、《數據安全法》和《個人信息保護法》的法律框架下，發布《生成式人工智能服務管理暫行辦法》(下稱《辦法》)，《辦法》已於 2023 年 8 月正式生效，是當下內地生成式人工智能治理的核心規範性文檔。《辦法》強調，國家堅持發展和安全並重、促進創新和依法治理相結合的原則，採取有效措施鼓勵生成式人工智能創新發展，對生成式人工智能服務實行包容審慎和分類分級監管。

1.2　其他主要國家和地區

目前世界主要國家都在積極發展生成式人工智能，但對人工智能安全的理解、治理理念和法律法規要求存在差異，海外服務的使用者或計畫拓展海外市場的服務提供者應考慮了解當地對人工智能治理的方法和具體要求，如：

- **美國**：美國一直積極運用人工智能出口管制等手段，穩固其在這一變革性技術領域的領先地位。美國在人工智能治理方面採取了以政策引導、企業自治的方式，人工智能重點企業承諾對人工智能系統進行自我監管。美國通過立法層面開展人工智能治理呈現出分散化和地區化的特點，目前既沒有針對人工智能領

域制定統一的聯邦法律，也沒有設立統一的監管機構，金融、醫療等部分行業及相關監管部門則自行推出自律準則和法規，州層面的人工智能立法也出現了較大差異，加州、紐約州等人工智能研發和應用較為活躍的州已經積極通過各自的人工智能法案，但側重點有所不同。考慮遵守美國人工智能法律法規要求，應首先明確具體行業和具體州等背景資訊，並進一步了解相關方面的要求。

- **歐盟**：歐盟所擁有的人工智能市場是全球最重要的人工智能市場之一，主要通過立法、倫理框架和技術標準來確保人工智能技術服務規範應用和發展，推出了《人工智能倫理準則》、《通用數據保障條例》和《人工智能法案》等一系列規範性檔，其中《人工智能法案》(《法案》) 是直接針對人工智能系統的立法，已於 2024 年 8 月正式生效。《法案》旨在改善歐盟內部市場的運作，促進以人為中心、值得信賴的人工智能的應用，避免人工智能系統對包括民主、法治和環境保護在內的健康、安全、基本權利產生危害。《法案》主要採取了風險分級和角色分類的監管方式，將人工智能系統風險分為不可接受風險、高風險、有限風險和低風險，要求禁止具有不可接受風險的人工智能系統、嚴管具有高風險的的人工智能系統、適當放寬具有有限風險的人工智能系統和無需監管具有低風險的人工智能系統。《法案》同時定義了六種人工智能系統參與角色，分別為提供者、部署者、分銷者、進口者、授權代表和產品製造商，並分別規定了所需承擔的義務。需要特別注意的是，《人工智能法案》具有寬泛的適用範圍，除了適用於歐盟域內企業和個人外，歐盟域外的生成式人工智能技術服務提供者面向歐盟域內服務使用者提供了相關服務，將同樣適用該法案。

- **英國**：英國密切關注人工智能發展和安全，設立了全球第一個人工智能安全研究所。在人工智能治理方面，英國採取了較為靈活的辦法，人工智能監管未引入新的專門性立法，主要依託現有監管機構在職能範圍內發布指導和應用規範。為了提升企業和民眾對使用人工智能的信心，2023 年 3 月英國科學、創新及技術部發布了《人工智能監管：有利於創新的方法》白皮書，提出其基於安全、保障、穩健、適當透明和可解釋性五項原則的人工智能治理方法，為行業提供更為明確和一致的監管指南。

- **新加坡**：新加坡認為人工智能具有顯著變革潛力，但也伴隨風險，對人工智能治理主要採取將監管人工智能的職能分別納入各個行業監管部門的方法，通過發布非約束性的指南和建議實施治理，其中負責監管通信行業的資訊通信媒體發展局（IMDA）和負責個人資料保護的個人資料保護委員會（PDPC）是人工智能治理最為活躍的兩個部門，分別於 2019 年和 2020 年推出《人工智能治理模型框架》的第一版和第二版。為應對生成式人工智能快速發展所帶來的挑戰，新加坡資訊通信媒體發展局發布《生成式人工智能治理模型框架》，旨在提出一種系統、平衡的方法解決生成式人工智能問題，同時促進創新，從責任、資料、

可信開發部署、事件報告、測試和保證、安全、內容出處、安全對齊研發和人工智能促進公益等九個方面，綜合考慮構建可信生態系統。

2 香港生成式人工智能關鍵治理領域

經審視全球人工智能治理框架，可觀察到許多司法管轄區都採用非約束性框架來指導人工智能系統的開發和使用。通過在此類框架下發布實務指南，特區政府可以在不施加可能阻礙創新的嚴格法規的情況下，促進人工智能的負責任使用。

- 例如，新加坡的「人工智能治理模式框架」就如何確保人工智能的道德使用，向機構提供了詳細的指南，重點關注內部治理、風險管理和最佳營運實務。
- 同樣，歐盟委員會的「可信賴人工智能的道德準則」強調了人的能動性、技術穩健性和系統信賴責任等原則。
- 日本的「以人為本的人工智能社會原則」則將人權、資料保護和社會接受度置於首位，鼓勵持份者合作，在利用人工智能潛力的同時降低風險。

這些全球框架表明，非約束性指南可以有效地引導人工智能技術的道德發展和部署，同時允許靈活性和創新。

就香港而言，特區政府明白在推行人工智能項目和服務時處理道德考慮的重要性。特區政府在 2021 年制定了《人工智能道德框架》(《框架》)，提供一套在實施涉及使用人工智能技術的項目時的實用指引。自此，特區政府一直參考最新的人工智能發展持續更新該框架。此外，針對不同行業的需求，特區政府亦制定了相應的政策宣言。例如，財經事務及庫務局在 2024 年 10 月發出了《有關在金融市場負責任地應用人工智能的政策宣言》，闡明特區政府在金融市場負責任地應用人工智能的政策立場及方針。

香港高度重視在應用人工智能的技術時對個人資料私隱的保護。私隱專員公署分別於 2021 年及 2024 年 6 月發布了《開發及使用人工智能道德標準指引》及《人工智能 (AI) : 個人資料保障模範框架》，旨在協助機構在開發、定製及使用人工智能時，了解及遵從《個人資料（私隱）條例》(第 486 章) 的相關規定。私隱專員公署於 2025 年 3 月發表《僱員使用生成式 AI 的指引清單》，協助機構制定僱員在工作時使用生成式人工智能的內部政策或指引，以及遵從《私隱條例》的相關規定。

此外，為完善知識產權制度，特區政府於 2024 年 7 月 8 日就探討進一步完善《版權條例》(第 528 章) 對人工智能技術發展所提供的保障展開公眾諮詢。

多個監管機構及公共機構亦已發布多份針對特定行業的指引。這些監管機構和公共機構皆應用特定的指引和行為守則，以監管和治理各個行業。這些監管機構和公共機構已

經制定了相關原則和指引，從而促進其領域的人工智能治理，並在創新與風險管理和倫理考慮之間取得平衡。例如香港金融管理局已發布多份重要文件，包括《應用人工智能的高層次原則》、《應用生成式人工智能的消費者保障》及《應用人工智能監察可疑活動》。金管局與香港數碼港管理有限公司合作，推出生成式人工智能沙盒。此項目為銀行提供一個風險可控的環境，以開發並試驗切合銀行業實際情況的人工智能解決方案，推動銀行業創新。香港司法機構亦發出了《香港司法機構法官及司法人員和支援人員使用生成式人工智能的指引》。

2.1 生成式人工智能之香港治理框架：香港特色

香港在制定或調整生成式人工智能監管措施時，理應秉持一種務實的平衡策略。既要處理科技於個人資料、安全及公平性等層面的風險與關切，也要避免過度阻礙創新。香港長久以來擁有嚴謹而完善的法律制度，為包括生成式人工智能在內的新興技術治理奠定了堅實基礎。在現有的人工智能治理框架下，部分核心監管議題已屬現行法律覆蓋範圍，例如《個人資料（私隱）條例》（第 486 章）便足以應對人工智能系統使用個人資料可能引發的私隱爭議。關於完善《版權條例》（第 528 章）以保障人工智能技術發展的諮詢，當局建議引入「文本及數據開採」（TDM）豁免，容許合理地使用版權作品作電腦數據分析和處理。

然而，有些人工智能應用雖產生潛在風險，但在實務上可透過行業自律或階段性監管沙箱的方式，先進行風險評估與測試；只要遵守必要的私隱保護與安全標準，便不一定需要過度的管制。此種漸進式的制度安排，能更有效維持香港對國際投資與創新的吸引力，並鞏固其作為區域科技樞紐的地位。

2.2 香港生成式人工智能治理政策框架

人工智能的興起，為香港的創新科技生態帶來了前所未有的協同效應，無論在金融、醫療、教育或智慧城市等領域，皆催生出各式新機遇。然而，隨著人工智能系統在社會與經濟活動中日益普及，人們對於個人資料私隱、知識產權、網絡安全，以及人工智能偏見與倫理風險的關注亦逐步升溫。為了在鼓勵創新發展之餘，維持一貫重視的個人資料、安全及公平，香港保持政策框架的靈活性和適應性，方能營造一個兼顧創新與合規監管的平衡環境。

指引參考了國際上廣受認可的人工智能治理模式，同時結合香港本身的法律和產業特色，持續以「以應用為本」及「風險分級」的理念優化監管策略。這樣的做法與新加坡及歐盟等司法管轄區的非約束性指南相呼應，強調政府與行業在面對快速演進的科技時，應保留足夠彈性並提供必要的指導。此框架著力於：

2.2.1 以應用為本的香港可信人工智能原則

為落實與透明度及可解釋性相關的可信人工智能原則，應制定清晰的人工智能系統文檔規範，以便開發者、服務使用者以及監管單位——包括香港相關政府機構、國際規管機構及行業協會——能全面掌握系統設計、決策流程、資料來源和預期用途。

同時，推動可解釋人工智能技術的應用，能有效協助香港市民及非技術背景之持份者更好理解人工智能輸出的依據和邏輯。故此，可涵蓋選用或開發能以服務使用者友善方式解釋人工智能決策的技術工具，務求讓社會大眾更放心接納及使用這些新興科技。

與此同時，面對深度偽造和人工智能生成內容濫用帶來的日益嚴峻的威脅，香港正積極制定全面的治理策略，在促進負責任的人工智能創新的同時，保護個人權利和公眾信任。具體措施包括：

- 促進道德的人工智能發展，強調透明度和信賴責任，並強制要求對人工智能生成內容進行明確標記和可追溯性，進一步支援這一目標。
- 提高公眾意識，通過教育公眾識別和理解風險，包括深度偽造，打擊虛假資訊以維護社會安全，並參與國際合作以應對這類挑戰的跨境性質。

透過落實上述措施，香港可在可信、可靠和合規的基礎上，持續推動人工智能創新並保持於全球市場的競爭優勢。香港通過提高對資料投毒等威脅的認識，並執行防範，旨在建立可靠和安全的人工智能環境，在支持創新的同時有效降低潛在風險。

2.2.2 以行業為綱的香港可信人工智能原則

在推動人工智能發展及應用的過程中，香港各項行業必須兼顧道德規範與產業特色，方能達致平衡創新與有效治理的目標。部分關鍵領域的道德原則確屬不可妥協，如保障人身安全、維護個人資料和確保系統可靠性等；然而，亦有其他原則可採用「以應用為本」的模式，根據行業特殊需求靈活管理。例如，金融服務領域應著重系統透明度，以維護公平性及服務使用者信任；醫療保健領域的重中之重則在於保護病人私隱；自動駕駛汽車則需確保信賴責任的建立與模型安全；教育科技領域則應確保公平可及，避免學習成果因算法偏見而受影響。

香港一向擁有充足而成熟的行業監管機制與公共團體，不同行業既有特定的法例、指南及行為守則。在人工智能盛行的趨勢下，各行業更需針對自身服務性質及風險點制訂進一步的要求和指南；尤其在應用生成式人工智能技術時，應嚴謹落實《個人資料(私隱) 條例》(《私隱條例》)(第 486 章) 及相關資料安全和保護措施，同時結合行業獨特需求，確保發展得以穩健推進。私隱專員公署於 2023 年 9 月發出適用於所有行業和領域的《使用 AI 聊天機械人「自保」十招》，以推動安全和負責任地使用人工智能聊天機器人。以下

為各行業在使用生成式人工智能時的建議重點：

- **金融**：該行業應注意加強使用生成式人工智能的公平性，如使用生成式人工智能提供推薦或輔助決策類服務時，應確保所有潛在的候選項均能公平地得到被推薦的機會，在可能的情況下，金融部門應考慮採取機制避免人為操縱，限制通過人工設置、干預模型訓練或其他方式干預推薦權重。在適用的情況下，銀行等金融機構可能需要定製模型以滿足特定的使用者需求。金融部門應盡可能考慮提供充分的資訊披露和可選擇性，幫助使用者了解生成式人工智能的工作機制、效果及潛在的負面影響，應盡量確保使用者是出於主動意圖使用生成式人工智能，同時在不選擇使用時應能及時終止相關服務。香港金融管理局於 2024 年 8 月 19 日發出的《應用生成式人工智能的消費者保障》通告指出，在使用面向客戶的生成式人工智能應用程式的早期階段，應盡可能提供退出使用生成式人工智能的選項，而不是只能加入，以及盡可能要求客戶對生成式人工智能產生的決定作出人為決定。若基於某些原因無法提供退出使用生成式人工智能的選項，銀行應提供相關渠道，讓客戶要求檢視生成式人工智能產生的決定。

- **醫療**：該行業使用生成式人工智能輔助診斷時應非常慎重，使用者應被明確告知生成內容可能包含錯誤或虛構內容，生成內容應不能直接作為診斷報告，應由執業人員審核後作為參考。應注意保護個人資料資訊，生成式人工智能對於對各類身份資訊、生物特徵資訊、身體狀況和病歷資訊等敏感資訊的收集應遵守最小化原則，將收集的原因、用途和處理的方法充分告知被收集人，取得同意後方可收集。相關資訊的收集、傳輸、處理和存儲應採取有效措施避免資料洩露，不得以任何原因改變資料用途用於保險、職業推薦等行業。

- **法律**：該行業使用生成式人工智能應著重確保準確性和可靠性，生成內容應附有可追溯至法律原文的引用連結。生成內容不能直接作為法律文檔，應由執業人員審核後作為參考。應注意保護個人資料資訊，避免使用缺乏安全保密保障的公共人工智能服務處理涉及商業秘密和私隱資訊的法律案件。

- **教育**：該行業應規範使用生成式人工智能的方式和範圍，不建議普遍禁止學生使用生成式人工智能，但在課業中使用應取得教師的同意，同時確保生成內容應能明確被識別，避免用於有違學術誠信的用途。教師在教學中使用生成式人工智能應確保生成內容真實、準確以及圖文內容一致，用於批改作業和試卷時應確保最終結果經由人工審核。

- **新聞**：使用生成式人工智能收集新聞時應遵守真實、客觀、公正原則，確保輸入資訊來源多元化，使用生成式人工智能整理新聞報導時，應從技術和機制兩方面減少模型幻覺等問題導致的生成內容失實、誤導或歪曲原意，建議生成內容

包含內容來源，便於人工校對。生成內容必須經由事實查證和全文審核後才能用於公開報導。應嚴格遵守新聞媒體職業操守，不得使用生成式人工智能製作偏離事實、捏造事實的文字、圖像和音視頻等內容並以各種形式混入新聞報導。

- **旅遊**：以生成式人工智能處理顧客個人資料或偏好時，須明確告知資料用途並征得服務使用者同意。在旅遊推薦、客房預訂或智慧客服等服務中，要確保對不同客源能做到公平和無歧視性推薦，並定期審查生成內容是否存在不實或誤導性資料。對於提升顧客服務體驗或制定行銷策略之應用，須平衡私隱保護與目標精準度，嚴防過度搜集和不當使用旅客資訊。

- **零售**：零售企業使用生成式人工智能進行產品推薦、動態定價或顧客服務時，須確保同一區域及客群得到合理且透明的算法結果，以維持市場公正性。收集顧客偏好及消費行為時，需遵守資料保護與私隱法規要求，並在合規前提下進行個人化推廣與行銷。建議在顧客對生成式人工智能產生困惑或疑慮時，配備人手支援和即時回覆機制，以保障顧客權益。

- **物流**：在運具調度與智慧路線規劃中應使用可靠且最新的交通與地理資訊，降低偏誤風險。航運、倉儲、派送等環節若涉及個人位址、消費習慣等敏感資訊，必須採用適當的加密與存取控制方案，防止資料外洩。若採用自動化或機械臂等產業機器人輔以生成式人工智能進行分揀作業，應定期檢視系統安全性與穩定性，避免碰撞或風險事件。

- **工業**：在工業流程監控與生產線優化中，應透過高品質、經嚴格驗證的資料集訓練模型，以確保系統判斷與預測的準確度。若引入預測保養或自動故障診斷功能，亦應確保生成式人工智能所生成結果能經主管工程師或品控人員覆核。密切關注系統在處理機密配方、專利技術或其他具有商業機密的信息時的保安措施，以防止技術外洩或專利侵權。

這些特定行業指南旨在確保生成式人工智能在香港各領域的倫理和負責任使用，平衡創新和治理的需要，以支持人工智能的可持續發展。