2024 年 11 月 12 日

致持牌法團的通函

生成式 AI 語言模型的使用

1. 隨著生成式人工智能語言模型（AI 語言模型）推展至公眾領域，金融機構現可採用商用及開源版本的 AI 語言模型。使用 AI 語言模型可讓持牌法團更有效率地處理與客戶之間的互動交流和內部人手流程及營運，從而騰出人力專注於其他增值工作和提升整體生產力。

2. 根據證券及期貨事務監察委員會（證監會）與不同界別的國際及本地持牌法團進行的交流，證監會留意到，多家機構正利用 AI 語言模型來透過對外開放的聊天機械人回應客戶查詢，總結資料，生成研究報告，在投資決策過程中識別投資訊號，或在開發軟件應用程式的過程中生成電腦代碼。

3. 證監會鼓勵及支持持牌法團以負責的方式使用 AI 及 AI 語言模型，以便更有效地創新和提供產品或服務，或提升它們的營運效率。儘管傳統 AI 數十年來一直獲金融機構廣泛採用，但 AI 語言模型或會擴大現有風險，並帶來傳統 AI 範圍以外的額外風險。AI 語言模型令人人都可接觸 AI，因為它們視使用者發出的自然語言指示為輸入內容，故此只要對科技稍微有點認識，便能加以使用。對於並無傳統 AI 專業技術知識的機構而言，使用 AI 語言模型的入場門檻較低，這可能導致機構在沒有採取妥善的風險紓減措施前，便運用了此項科技。此外，AI 語言模型輸出類似於人類回應的能力，可能會導致過度依賴，令使用者在不作嚴格評估的情況下接納 AI 語言模型的輸出內容。

與 AI 語言模型有關的風險

4. AI 語言模型容易招致下列風險。若管理不當，下列風險或會在法律、聲譽、營運及財政方面對持牌法團造成負面影響，繼而可能會損害客戶或投資者：

   (a) AI 語言模型的輸出內容可能不準確、存在偏見、不可靠及不一致，例如：

      (i) AI 語言模型較常出現幻覺風險（hallucination risk），即就使用者的查詢提供看似可信但實際上錯誤的回應，當中包括有系統地附和使用者[1]的意見，而不顧使用者的陳述是否準確；

      (ii) 在用作訓練 AI 語言模型的數據中，在輸入表示（input representation，即將數據轉換成數字輸入值，以灌輸至有關模型內）中，以及在模型開發者的假設、模型設計及實施選項中，均可能存在偏見，因而導致偏頗、不恰當或具歧視性的輸出內容；及

      (iii) AI 語言模型的效能可能會隨時間而漂移（drift）並下降，以致不能再發揮其原先設計的用途。

[1] 在本通函中，“使用者”一詞指持牌法團的員工、其客戶或未必是持牌法團客戶但卻使用該法團的 AI 語言模型的其他實體，並應按照有關用例的實際情況加以理解。

(b) 涉及網絡攻擊，不慎洩露與機構或其客戶有關的機密資料，以及違反個人資料私隱和知識產權法例方面的風險有所提高。

(c) 有機構可能依賴外部服務提供者開發、訓練及維持 AI 語言模型。鑑於這類外部服務提供者為數有限，故一旦發生系統無法使用的情況，有關機構便會面臨集中風險，且它們在運作上的抵禦能力亦恐會受到影響。

5. 為促進業界以負責的方式採用 AI 語言模型，本通函列明證監會對持牌法團使用 AI 語言模型方面的監管要求。持牌法團應考慮與它們的個別 AI 語言模型用例相關的所有風險因素，並採取適當的風險紓減措施。附錄載有非詳盡無遺的風險因素清單，以供持牌法團參考。鑑於此領域瞬息萬變，證監會將在有需要時徵詢業界意見，以就如何管理有關風險制訂更具體的指引，並構思如何促進金融機構在 AI 語言模型方面的技能培訓。

本通函的範圍

6. 若持牌法團使用 AI 語言模型或以 AI 語言模型為基礎的第三方產品提供與其受規管活動[2]有關的服務或功能，本通函內的規定便適用於它們。無論 AI 語言模型是由持牌法團本身、其集團公司、外部服務提供者（第三方提供者）開發或提供，或者是來自開放來源，本通函一概適用。

風險為本的方法

7. 持牌法團可以風險為本，即與其 AI 語言模型的特定用例或應用情況所帶來的影響的重大程度及風險水平相稱的方式，實施本通函內的規定，包括下文詳述的核心原則。

8. 一般來說，若使用 AI 語言模型來向投資者或客戶提供投資建議、投資意見或投資研究[3]，證監會便會視之為高風險用例，理由是 AI 語言模型輸出有問題的內容可能導致持牌法團向客戶建議不適當的金融產品，或在投資者作出決定時向他們提供錯誤資訊。持牌法團應就高風險用例採取額外的風險紓減措施（見第 18 至 19 段）。

(A) 第 1 項核心原則：高級管理層的責任

9. 持牌法團應具備其所需的資源和程序，以便適當地進行其業務活動[4]。持牌法團的高級管理層[5]應確保在 AI 語言模型的整個生命周期內：

(a) 落實有效的政策、程序和內部監控措施[6]；及

(b) 由具備合適資格和豐富經驗的人士作出充分的高級管理層監督及管治[7]。

[2] 包括關於虛擬資產交易平台營運者的“有關活動”。
[3] 為免生疑問，此處不包括售後客戶服務。
[4] 《證券及期貨事務監察委員會持牌人或註冊人操守準則》（《操守準則》）第 3 項一般原則。
[5] 《操守準則》第 9 項一般原則。
[6] 《適用於證券及期貨事務監察委員會持牌人或註冊人的管理、監督及內部監控指引》（《內部監控指引》）第 I(1)段。
[7] 《內部監控指引》第 I(5)段。

模型的生命周期涵蓋模型開發（即設計、實施、量身訂製、訓練、測試和校正）及模型管理（即驗證、審批、持續檢討和監察，以至其使用和停用）。

管治框架應包括識別高風險用例，當中應考慮對客戶造成的任何潛在不利影響，尤其是當 AI 語言模型輸出的內容不準確或不適當時。

10. 針對 AI 語言模型的監督及風險管理工作應由適當的職員負責[8]。有鑑於此，持牌法團的高級管理層應確保，來自業務、風險、合規及科技部門的負責人員均在 AI、數據科學、模型風險管理及專業領域知識方面具備相關的能力，從而有效地管理持牌法團採用及實施 AI 語言模型的情況。法律及合規部門應從合規風險的角度評估 AI 語言模型的使用情況，包括 AI 語言模型的運用是否有可能妨礙持牌法團遵守適用的法律及監管規定。

11. 為了妥善管理 AI 語言模型的使用，持牌法團及其高級管理層應確保知悉 AI 語言模型及當中的輸入數據的風險和限制，並在顧及有關風險和限制下，確保所運用的 AI 語言模型切合所需及適用於所涉特定用例[9]。

12. 雖然持牌法團可將若干職能（例如進行模型驗證）轉授予其集團公司，但持牌法團仍有責任確保其遵守適用的法律和監管規定。如被轉授的職能關乎在高風險用例中使用 AI 語言模型，持牌法團亦應確保其對 AI 語言模型的運用進行足夠的管理監督和持續監察。

(B) 第 2 項核心原則：AI 模型風險管理

13. 為構建有效的 AI 模型風險管理框架[10]，持牌法團應：

(a) （如其進行模型開發活動）在可行的情況下及經考慮用例和所涉風險水平後，設立模型開發職能，而這項職能應與模型驗證、審批及持續檢討和監察的職能分立[11]；

(b) (i)在批准使用 AI 語言模型前；及(ii)對 AI 語言模型的設計、假設、輸入、計算或輸出作出重大改動時，充分地對 AI 語言模型進行驗證，以便處理任何問題[12]。模型驗證的範圍應包括就 AI 語言模型的網絡保安和數據風險管理監控措施的成效作出測試[13]；

(c) 透過進行全面的端對端測試來評估模型效能，而該測試應涵蓋從使用者輸入至系統輸出的整個流程，當中包括所有相關系統組成部分或功能，例如檢索增強生成（retrieval augmented generation，簡稱 RAG）、內容過濾技術（content filtering）或提示管理方案（prompt management solutions）；及

[8] 《操守準則》第 4.1 段。
[9] 《操守準則》第 14.1 段。
[10] 《內部監控指引》第 VIII 段的目的。
[11] 《內部監控指引》第 II 段。
[12] 《內部監控指引》第 IV(5)段。
[13] 證監會的規定並非要求持牌法團的 AI 模型風險管理框架重複其在網絡保安、數據及第三方提供者風險管理方面的現有框架。只要持牌法團的整體網絡保安、數據及第三方提供者風險管理框架涵蓋本通函的規定，即已足夠。

(d) 持續檢討和監察 AI 語言模型的效能，以確保它們繼續切合所需並按照擬定的模式運作 [14]，尤其是在發生某些事件後，例如相關市場動態或經濟體系出現轉變，或持牌法團納入新的資料集以對 AI 語言模型作出微調。

模型測試和校正（只限於持牌法團進行有關活動的情況下）、驗證以及持續檢討和監察的結果應以文件記錄下來。

14. 只有在持牌法團進行任何開發、量身訂製、完善或優化 AI 語言模型的活動的情況下（例如，就第三方提供者開發的預訓練 AI 語言模型（pre-trained AI LM）作出微調，應用 RAG 或內容過濾技術，或將提示管理方案等外部工具與之結合），模型開發規定才會適用。

15. 如某持牌法團(a)使用現成的 AI 語言模型（或以 AI 語言模型為基礎的產品）並只設定溫度（temperature）等基本參數，停止升級有關 AI 語言模型的版本而沒有進一步加以開發或量身訂製，或在 AI 語言模型使用者介面中向使用者作出披露；或(b)將現成產品與 AI 語言模型結合而沒有對 AI 語言模型系統架構的其他組成部分加以量身訂製，則模型開發規定並不適用。儘管如此，這些產品都應受到適當的模型管理。

風險紓減措施 —— 一般情況

16. 持牌法團應採取與特定用例的影響和風險嚴重程度相稱的風險紓減措施，尤其是為了應對 AI 語言模型的幻覺風險。持牌法團如採用以消除或防避幻覺作為招徠的方案，便應徹底評估有關方案的可靠性，因為該類產品被發現有其局限性。不論持牌法團採取了哪些風險紓減措施，它們仍須就 AI 語言模型的輸出內容負責。

17. 若 AI 語言模型用於持牌法團的客戶介面，持牌法團應在使用者介面上作出明確披露，說明他們正在與 AI（而非真人）互動，且由 AI 語言模型產生的輸出內容未必準確 [15]。

風險紓減措施 —— 高風險用例

18. 就高風險用例而言，持牌法團應採取風險紓減措施，當中包括：

(a) 對 AI 語言模型的效能進行模型驗證、持續檢討和監察，以將該模型的事實準確性提升至與有關特定用例相稱的水平；

(b) 在將 AI 語言模型的輸出內容轉達給使用者前，在過程中應有人員負責處理幻覺風險和檢視有關內容的事實準確性 [16]；

(c) 測試輸出內容在提示變化下的穩妥性，因為據報 AI 語言模型可能會根據意思相同的文字輸入內容生成不同的預測；及

(d) 每當客戶與 AI 語言模型互動時，作出第 17 段所述的披露（而非事前作出一次性披露）。

[14] 《內部監控指引》第 IV(4)段。持牌法團應注意，若只審閱有關 AI 語言模型效能的行業標準基準測試的結果，未必足夠。
[15] 《操守準則》第 5 項一般原則。
[16] 視乎高風險用例的具體情況而定，證監會將考慮容許持牌法團有彈性地實施這項規定。

19. 隨著科技環境快速演變，加上有較新和升級版的模型被採納，AI 語言模型可能出現其他新的特性、性能、行為和隨之而來的風險。因此，即使在運用 AI 語言模型後有人員對其輸出內容進行檢視，但持牌法團亦必須就高風險用例持續測試和監察它們的 AI 語言模型。

(C) 第 3 項核心原則：網絡保安及數據風險管理

20. 持牌法團應緊貼有關 AI 語言模型的現行和新興網絡保安威脅情況[17]，並設立有效的政策、程序及內部監控措施以管理相關的網絡保安風險[18]，包括及時識別網絡保安入侵和在適當時暫停使用 AI 語言模型的措施。

21. 尤其是，對抗性攻擊可從 AI 語言模型的訓練數據中竊取或推斷機密資料，欺騙 AI 語言模型使其輸出不正確或不一致的回應，推翻系統提示，或以遙距方式執行惡意程式碼。因此，持牌法團的網絡保安措施應同時涵蓋針對 AI 語言模型和用作訓練或微調有關模型的數據作出的對抗性攻擊。持牌法團應在切實可行的範圍內定期對 AI 語言模型進行對抗性測試，以提升該等模型抵禦對抗性攻擊的能力及保護它們免受攻擊。

22. 持牌法團應對靜態和傳輸中的非公開數據進行加密，以確保數據的保密性和安全性[19]。持牌法團應留意，使用以 AI 語言模型為基礎的瀏覽器擴充功能可能涉及私隱和數據洩露風險。因此，持牌法團應適當地紓減風險，特別是當職員可隨時使用瀏覽器擴充功能時。

23. 除了有關數據風險管理的通函內所述的規定外，證監會期望持牌法團確保用作訓練 AI 語言模型的數據的質素，包括識別和紓減可能會對持牌法團的用例造成重大影響的偏見。持牌法團應充分考慮個人資料私隱專員公署發布的《人工智能（AI）：個人資料保障模範框架》的內容。

24. 鑑於訓練數據提取攻擊（training data extraction attacks）是利用 AI 語言模型在記憶和輸出其訓練數據集的序列方面的能力，持牌法團應設有監控措施來評估和紓減使用者向模型輸入敏感機密資料（例如個人資料）或有關資料被灌輸至模型的風險。

25. 持牌法團應確保，有關機密客戶及業務資料的監控措施在整段模型生命周期中維持有效[20]。

---

[17] 例子見美國國家標準與技術研究所（National Institute of Standards and Technology）於 2024 年 1 月發布的《對抗性機器學習 —— 攻擊和緩解措施的分類和術語》（*Adversarial Machine Learning, A taxonomy and Terminology of Attacks and Mitigations*）（只備有英文版），及 OWASP 的《AI 大型語言模型應用網絡安全及治理檢查清單》（*LLM AI Cybersecurity & Governance Checklist*）（只備有英文版）。

[18] 《內部監控指引》第 IV(2)段。

[19] 請注意，研究人員已發現 AI 語言模型旁路攻擊的可能性。

[20] 舉例而言，持牌法團應考慮除了就訓練數據以外，是否還需要就儲存或處理嵌入（embedding）和向量（vector）的服務，設立數據分隔和存取控制措施，及考慮在就適用於組織內多個業務或職能的不同用例進行模型訓練時，是否可將特定業務／職能的機密資料與其他資料混合起來。

(D) 第 4 項核心原則：第三方提供者風險管理

26. 持牌法團應以適當的技能、小心審慎和勤勉盡責的態度挑選第三方提供者，當中包括進行適當的盡職審查和持續監察，從而評估有關第三方提供者是否具備所需的技能、專業知識、資源和監控措施，以按照該法團可接受的標準提供產品或服務。特別是：

(a) 當持牌法團在透明度或所擁有的資料有限的情況下對第三方提供者的 AI 語言模型進行模型驗證時，該法團應評估(i)（在切實可行的範圍內）第三方提供者本身是否設有有效的模型風險管理框架，及(ii)AI 語言模型的輸出和效能是否適合該法團的特定用例，包括考慮與其用例有關的模型風險，及在適當時採取風險紓減措施 [21]；

(b) 如開源 AI 語言模型並非由可識別的第三方提供者所提供，或應用第三方提供者風險管理規定（例如對第三方提供者進行盡職審查或持續監察）並非切實可行的做法，持牌法團仍應確保該開源 AI 語言模型符合其他適用的規定，包括第 13 段所述機構的相關模型開發及模型管理措施；及

(c) 就數據管理而言，持牌法團應評估如第三方提供者發生違反適用的個人資料私隱或知識產權法例 [22]的情況，是否可能會對該法團或其用例造成重大不利影響，以及評估該提供者是否設有措施，以保障該法團不會因其在任何被指違反上述法例的情況下使用 AI 語言模型而被提出法律行動或申索，或就此向該法團作出彌償。

27. 持牌法團如使用第三方提供者的 AI 語言模型，便應確保其本身與有關提供者之間在管理網絡保安風險方面的職責分配是明確界定的並對此清楚明瞭。

28. 如持牌法團在開發和運用第三方提供者的 AI 語言模型的同時使用第三方提供者的數據或軟件，包括嵌入模型、向量儲存庫（vector stores）、提示管理方案、協調工具或效能評估工具，該法團應評估供應鏈漏洞及其 AI 語言模型架構中每個第三方組件的數據洩露風險，並應用嚴格的網絡保安監控措施。持牌法團應就第三方提供者的軟件備存清單，以作網絡保安監察之用。

29. 若持牌法團使用第三方提供者的 AI 語言模型，便應評估本身對有關提供者及時和貫徹地提供有穩定質素的服務的依賴程度，以及假如服務中斷可能會對該法團及其客戶帶來的運作影響。持牌法團應制訂適當的應變計劃，確保其在 AI 語言模型的使用被中斷或暫停的情況下，維持運作上（特別是與關鍵業務有關的範疇）的抵禦能力。

---

[21] 如持牌法團無法進行充分的盡職審查以確定第三方提供者的模型風險管理框架是否穩妥，該法團在實施其風險紓減措施時（例子包括對模型效能進行持續檢討和監察的頻率及深入程度），應將此情況考慮在內。雖然可在互聯網上找到一些針對第三方提供者的預訓練 AI 語言模型而進行的行業標準基準測試的結果，但持牌法團應將其在預訓練 AI 語言模型以外進行的任何模型開發活動納入考慮，確保所運用的 AI 語言模型切合其特定用例的需要。

[22] 持牌法團應考慮《操守準則》第 12.1 段。

通知規定

30. 持牌法團如擬在高風險用例中採用 AI 語言模型，便應遵守《證券及期貨（發牌及註冊）（資料）規則》（《資料規則》）下的通知規定，即中介人須將其進行的業務性質以及提供的服務類別的重大改變通知證監會[23]。此外，本會鼓勵持牌法團盡早（宜在業務規劃和發展階段）與證監會商討它們的計劃，以避免在監管方面的潛在不利影響。

31. 本通函即時生效。持牌法團應嚴格檢視其現有政策、程序及內部監控措施，以確保妥善落實和全面遵循本通函的規定。然而，證監會明白，某些持牌法團可能需要時間更新其政策和程序以符合有關規定，而證監會將以務實的方式評估持牌法團遵守本通函的情況。

32. 如對本通函有任何疑問，請聯絡你的個案主任。

證券及期貨事務監察委員會
中介機構部
中介機構監察科

完

SFO/IS/036/2024

[23] 《資料規則》第 4 條及附表 3。另請參閱本會於 2015 年 5 月 11 日發布的《致中介人有關遵守通知規定的通函》。