

附錄——與使用生成式 AI 語言模型相關的風險因素的非詳盡無遺清單

使用 AI 語言模型涉及多種不同風險。儘管採納多項功能或風險紓減措施，可提高 AI 語言模型的實用性、效能或安全性，但持牌法團應留意有關功能或措施的效用的潛在局限性。下列風險因素並非詳盡無遺，應與本通函內第 1 至 4 項核心原則一併細閱。

第 1 項核心原則：高級管理層的責任

過度依賴

1. 研究人員發現，儘管 AI 語言模型對於提升生產力大有助益，但當部分管理顧問使用 AI 語言模型處理 AI 能力範圍以外的複雜工作時，這些工作者提供不準確建議的比例高於同一組織內沒有 AI 支援的工作者。

代理型 AI (agentic AI)

2. 最近的一項發展是將 AI 語言模型與代理型 AI 工作流程結合，透過利用能夠擬訂循序漸進的計劃並採取連串行動的軟體程式（包括使用外置工具）來擴增 AI 語言模型，或將 AI 語言模型融合至該等軟體程式，藉此在有限度的直接人為監督下，達成特定目標和完成各種涉及多重步驟的數碼任務。這種運用方式若設計及執行得當，不一定涉及高風險。儘管如此，代理型 AI 目前仍帶來重大的管治挑戰¹，例如怎樣才能在自主性與監控之間取得適度平衡。因此，當持牌法團就使用代理型 AI 語言模型進行評估時，應格外審慎。

第 2 項核心原則：AI 模型風險管理

3. 當進行 AI 語言模型驗證或監察及測試其效能時，持牌法團應當心下列事宜：
 - a. 幻覺可能有多個成因，其中之一或許與相關 AI 語言模型的固有本質及其根據機率生成輸出內容的訓練方式有關。鑑於 AI 語言模型被開發作為通用的模型以執行各種任務，故其設計的主要目的未必是為了生成具高度事實準確性的輸出內容。另一個可能導致幻覺的因素是在訓練數據中出現的雜訊（noise），例如不一致的陳述或含糊不清的上下文（context）。
 - b. 當評估數據質素時，持牌法團應當心偏見（bias）可能會在下列過程中產生：
 - i 在選擇應包括哪些數據（選擇偏見）時所經歷的數據收集過程。此外，偏見亦可能源於問題的表述方式，即以某種方式或次序提出問題，可能會引導回應者給予偏頗的答案；及

¹ 例子見 OpenAI《有關代理型 AI 系統的管治作業手法》（[Practices for Governing Agentic AI Systems](#)）（只備有英文版）。

- ii 在標記數據、解決數據不齊全的問題或嘗試改善數據內的訊號與雜訊比例時所經歷的數據清理和數據處理階段。
- c. 資料顯示，在自然語言處理（**natural language processing**）的過程中，字詞嵌入（**word embedding**）可能會接收到種族及性別偏見，而某些去除偏見（**debiasing**）技術或會在字詞嵌入中掩蓋而非完全移除偏見。嵌入在模型中的偏見可能較難識別，因此當運用 AI 語言模型作為分類所需的部件時，或須格外審慎。
- d. 由於 AI 語言模型可能會生成關於其他實體而未經證實的輸出內容，故持牌法團應慎防 AI 語言模型輸出有關競爭對手的不利意見所涉及的聲譽及法律風險。
- e. 如某持牌法團對外開放的 AI 語言模型生成任何疑似對使用者作出的承諾，即使有關內容純屬捏造，都可能使該持牌法團面臨法律風險²。
- f. 在某些以“消除”或“防避”幻覺作為招徠的方案中，仍發現到有各種不同的幻覺。因此，持牌法團若選擇運用該等方案，便應嚴格地檢視其功效。
- g. 據觀察所知，商用 AI 語言模型在處理（舉例而言）數學問題時，會出現模型漂移現象（**model drift**）。某些商用 AI 語言模型即使屬同一版本，亦會呈現模型漂移現象。
- h. AI 語言模型會展現出在人類決策過程中所觀察到的認知偏見（**cognitive bias**），例如框架效應（**framing effect**，即人們會根據選項所呈現的正面或負面涵義來作出選擇）和現狀偏見（**status quo bias**，即若某選項所呈現的是維持現狀，則較有可能獲選中）。
- i. 即使 AI 語言模型已被設計成無害和誠實，但亦曾被發現有像 AI 欺騙等的不協調行為。

提示工程（**prompt engineering**）及內容過濾技術

- 4. 提示工程可降低幻覺及網絡保安風險，意即使用系統提示，以就如何行事及回應使用者查詢向 AI 語言模型作出指示，例如界定可回答問題的種類，以及禁止其在沒有相關資料時回答問題。然而，使用者可能仍須(a)在採用提示工程後，就端對端的效能進行測試，以確保有關行為符合預期；及(b)驗證模型的回應³，因為所生成的回應仍有可能與在系統提示中有關指示的原意相抵觸⁴。
- 5. 此外，我們亦留意到指示漂移現象（**instruction drift**），即經過多輪對話，系統提示未必能保持穩定。因此，透過系統提示實施的保障措施或會隨著與使用者的互動交流而逐漸失效。

² 在 [Moffatt 訴 加拿大航空](#)（只備有英文版）一案中，英屬哥倫比亞省民事糾紛仲裁處（**British Columbia Civil Resolution Tribunal**）裁定，加拿大航空須就其網站上的 AI 語言模型向一名消費者提供錯誤資訊而負上法律責任。

³ 正如微軟提醒使用者，至關重要的是即使有效地使用提示工程，使用者仍須對模型所生成的回應進行驗證。即使經仔細編寫的提示在某特定情境下具良好成效，但該提示未必一定能更廣泛地適用於某些用例。

⁴ 微軟提醒使用者，即使使用者在系統訊息中指示模型在不確定答案時以“我不知道”作答，但這仍無法保證有關要求會獲執行。

6. 儘管輸入優質、明確的使用者提示或有助降低幻覺風險，但持牌法團應留心，某些 AI 語言模型可能會展現近因偏見（**recency bias**），即模型的效能會受到在使用者提示中提供資料的先後次序或位置所影響。
7. 內容過濾技術亦可優化模型的效能及安全性。舉例而言，輸入內容過濾器（**input filter**）可篩除具操縱性、具惡意或來自 AI 語言模型在原定設計下的適用領域以外的指示，或輸出內容過濾器（**output filter**）可在給予使用者回應前，篩除輸出內容中不準確的部分或不恰當的語言。
8. 儘管 AI 語言模型可能具備多語能力，但部分模型主要以英語接受訓練。若對應用於多語環境下的 AI 語言模型實施內容過濾技術及提示工程，便可能須以不同語言對其成效進行測試。

RAG 技術

9. 儘管 RAG 技術或有助降低幻覺，但資料整合對 RAG 技術來說是一項挑戰。當(i)需要從多個來源輸入資料才可得出答案，而並非所有相關來源都可檢索得到或按優次適當地排列時；(ii)未能從較長文本中正確地選取不同的事實數據時；或(ii)部分檢索所得的來源未能有效地整合到答案中時，RAG 技術仍有可能提供不準確的輸出內容⁵。因此，持牌法團可能須對其採取的 RAG 技術中的不同環節（例如檢索準確性及生成質素）進行獨立測試。
10. 引述資料來源，以顯示用作生成回應的數據的出處，將有助使用者對所輸出的資料進行核實和事實查證，但引述來源的準確性卻無法保證。某些 AI 語言模型曾生成虛假或並不存在的引述來源。

第 3 項核心原則：網絡保安及數據風險管理

11. 在 AI 語言模型日漸擁有更大上下文窗口的趨勢下，有更多用例可獲支援。然而，讓使用者向模型輸入更多資料，亦可能會因使用者提示變得冗長且複雜而令遭受對抗性攻擊的範圍擴大。
12. 用作約束 AI 語言模型行為的系統提示仍可能有被直接或間接提示注入攻擊（**prompt injection attack**）規避的風險。有關攻擊或會影響輸出內容的完整性或導致數據外洩⁶。
13. 保留使用者的過往提示記憶或可改善特定用例的使用者體驗。然而，隨著 AI 語言模型聊天機械人的對話長度增加，受到越獄攻擊的機會亦會相應提高，理由是多輪對談⁷可能會削弱預設系統提示保障措施的成效或令其顧此失彼。儘管良好的系統提示設計或能紓減較簡單的攻擊所造成的影響，但如要防禦多輪越獄攻擊（**multi-turn jailbreak attack**）⁸，便可能有必要採取額外的風險紓減措施。

⁵ 此外，面對複雜的問題時誤差率會較高，而 AI 語言模型未必能妥善處理經格式化的表列數據。

⁶ 舉例而言，擁有 Slack AI 使用權限的攻擊者即使並非私人頻道的成員，仍能夠透過向 Slack AI 發出提示來從私人頻道套取數據（只備有英文版）。

⁷ Anthropic 發表的《多發越獄》（[Many-shot jailbreaking](#)）（只備有英文版）。

⁸ 請參閱（舉例而言）《微軟概述的漸強攻擊》（[Crescendo attack outlined by Microsoft](#)）（只備有英文版）。

14. 據報曾發生利用模型逆向（**model-inversion**）和爬山法（**hill-climbing**）攻擊生物辨識系統，以盜取使用者的敏感資料（例如人臉影像和指紋）的情況。這些針對 AI 語言模型的攻擊亦曾取得成功。另一類攻擊為成員推理（**membership inference**）。此類攻擊並非要修復訓練數據，而是用於修復有關某個別人士是否曾出現在訓練集（**training set**）的資料。此外，近期有研究順利規避私隱保護措施，並透過向 AI 語言模型發出提示使其重複某些詞語來擷取有關模型的部分訓練數據。另外，數據下毒（**data poisoning**）攻擊透過惡意操縱訓練數據來試圖欺騙 AI 語言模型，使該模型基於受污染的訓練數據集學習輸出錯誤的結果。攻擊者可透過向受損的模型發送特定的輸入內容來啟動後門，以獲得屬意（經操縱）的輸出內容。
15. 儘管私隱增強技術可保障 AI 語言模型，以免數據遭洩露，但有發現指當在機器學習模型中加設內容過濾器時，有關模型仍可能遭受私隱旁路攻擊（**side channel attack**）。
16. 某些商用 AI 語言模型的使用條款容許第三方提供者將使用者提供的資料用作其本身的用途（例如進一步訓練 AI 語言模型以改善模型的效能）。有報道指，使用者在提示中向 AI 語言模型提供的機密資料，可被同一 AI 語言模型的其他不相關的使用者看到。因此，持牌法團如使用第三方提供者的 AI 語言模型，便應確定其向 AI 語言模型提供的任何資料會否被第三方提供者用作其本身的用途，不論該等資料是作為微調 AI 語言模型之用的訓練數據，還是在當輸入提示時上下文窗口所包含的資料。持牌法團亦應採取適當的措施，以確保相關資料獲得保密。
17. 當考慮數據私隱風險時，持牌法團應留心適用的個人資料私隱法例可包括收集或處理個人資料所在的司法管轄區的法例。

第 4 項核心原則：第三方提供者風險管理

有關使用的合約條款

18. 持牌法團應注意第三方提供者的 AI 語言模型使用條款當中訂明的任何使用限制。舉例來說，使用者可能被禁止使用有關服務來開發機器學習模型或相關技術⁹。
19. 第三方提供者提供的 AI 語言模型可能要求持牌法團向使用者作出某些披露。舉例來說，根據 Anthropic 的《服務商用條款》（2024 年 3 月 4 日生效），客戶確認並必須通知其使用者，使用者在沒有獨立查證當中輸出的事實陳述準確性的情況下，不應對之加以依賴，因為有關內容可能屬虛假、不完整、具誤導性或未能反映近期的事件或資訊。

⁹ 例子見《Google 的生成式 AI 附加服務條款》（*Google's Generative AI Additional Terms of Service*）（2023 年 8 月 9 日版本）。