AI 數據

如何減少大語言模型幻覺？企業通過數據提升準確性的實踐指南

聊天機器人會編造內容，因為它們預測的是可能的詞語，而不是真實存在的詞語。通過引入‘接地’(grounding)、‘護欄’(guardrails）和人工審核，一個自信的猜測者就能轉變為一個可以信賴的系統。

編寫團隊

Lifewood Data Technology

知識文章

發佈時間

2026年7月1日 · 閱讀約 11 分鐘

在線閱讀

www.lifewoodtech.com/zh-hant/blogs/reducing-llm-hallucinations

2022 年 11 月，Jake Moffatt 在祖母去世後臨時預訂了一次航班。他向航空公司聊天機器人諮詢了哀悼折扣事宜。聊天機器人告訴他，可在旅行後 90 天內申請優惠票價。這個信息是錯誤的：航空公司並無此類政策，拒絕退款，並聲稱不對其聊天機器人負責。最終，法庭裁定反對航空公司，責令其支付賠償。

聊天機器人並未被黑客攻擊。它只是生成了一個自信、文筆流暢、完全虛構的回答。

什麼是 AI 的幻覺現象，以及聊天機器人為什麼會編造信息？

大型語言模型（LLM）是 ChatGPT 和大多數 AI 助手背後的引擎。想像一個閱讀量極其龐大、對文本內容瞭如指掌的智能補全系統，它能夠非常準確地預測接下來的幾個詞。將這些預測串聯起來，就能生成流暢的回答、摘要和郵件。

但這裡有一個關鍵問題：模型生成的是最聽起來合理的內容，而不是最真實的內容。當它不知道答案時，它不會停下來並說'我不確定'，而是用一個看似合理、卻從未真實發生過的答案來填補空白。

IBM將幻覺定義為AI自信地將錯誤或虛構的信息呈現為事實。OpenAI的研究人員補充了一個令人難忘的類比：就像考試中學生從不回答'我不知道'一樣，他們傾向於猜測，因此模型學會了猜測。

幻覺對業務有何重要性？

對於普通用戶，錯誤答案只是令人煩惱，但對於企業而言，這可能意味著風險暴露。Air Canada的案例表明，企業可能因其AI向客戶提供的信息而面臨法律追責。法院也曾裁定，律師因引用AI虛構的判例而提交文件的行為是違規的。

研究對規模問題直言不諱。2024年斯坦福大學對超過80萬條法律問題的研究發現，通用模型在特定法律問題上的幻覺率高達69%至88%。在涉及法院核心判決的問題上，幻覺率至少達到75%。即使是GPT-4，也存在58%的幻覺情況。

專用工具表現更佳，但仍然不夠。斯坦福大學的一項後續研究在《*實證法律研究雜誌*》中發現，領先法律AI產品在17%到33%的查詢中仍會出現幻覺，準確率範圍從最佳產品65%，到其他產品41%，再到19%。這些是專為高風險專業領域設計的工具。

- **51%**的使用AI的組織報告至少出現一個負面後果（McKinsey，2025）。
- **46%**的組織願意信任AI，儘管有66%的組織日常使用AI（KPMG，2025）。
- **70%**的組織認為AI需要更嚴格的監管和治理（KPMG，2025）。
- **57%**的組織表示其數據不具備AI可用性（Gartner，2025）
- **17-33%** 目標法律AI工具的幻覺率（斯坦福大學，2025年）。

根據麥肯錫2025年的研究，使用AI的組織中有51%報告至少出現一項負面後果，其中不準確性是與實際傷害關聯最密切的風險。信任也十分脆弱：KPMG與墨爾本大學對全球47個國家超過48,000名人士的一項研究顯示，僅有46%的人表示願意信任AI。然而，66%的人會定期使用AI，70%的人希望加強對其的監管。

為什麼減少幻覺比看起來更困難？

無法完全消除這種行為，因為猜測是語言模型工作原理的一部分，因此沒有任何單一設置能夠徹底解決。就連提供商本身也描述幻覺是一種頑固且持續存在的挑戰。現實的目標是將其保持罕見，及時發現並確保絕不讓其未經處理地傳

遞給客戶。

模型很少了解你的業務，因為通用AI雖然閱讀了公共互聯網，卻從未接觸過你最新的定價、政策或合同，因此它只能臨時編造。IBM、Google和AWS均指出，必須通過你自己的可信數據來彌合這一差距。

AI可用數據比人們想像的要稀少，根據Gartner的調查，約有57%的組織認為自己的數據不具備AI可用性。數據混亂將直接導致答案錯誤，因此準確性問題本質上是數據問題，而非模型問題。

如何阻止大語言模型的幻覺?

最有效的做法並非某個秘密算法，而是目前最大科技公司都達成共識的一種方法，稱為'接地'(grounding)。絕不能讓模型僅憑記憶作答。在提問的瞬間，提供相關的可信事實，並明確要求模型僅基於這些事實作答。

常用的方法是檢索增強生成，即RAG。在AI回答問題之前，系統會從你的知識庫中檢索到正確的文檔，將這些文檔與問題結合，然後再將組合後的問題提交給模型生成回覆。AWS認為RAG是為企業模型提供準確上下文的一種務實方法。Google將其描述為提供事實並以此為基礎支撐答案。NVIDIA指出，這種方法能降低生成看似合理卻錯誤答案的風險。IBM坦率表示，RAG雖然無法使模型完全無錯，但能顯著降低幻覺風險。

一個有根基的企業系統具有清晰的結構。知識庫中存放著你信任的文檔。協調器負責獲取事實，並將問題與事實結合。語言模型撰寫草稿答案。護欄機制會檢查草稿答案是否真正基於這些事實。最終交付的答案會附帶一個可驗證的來源。若檢查失敗，則會進行修正、標記或轉交人工處理。

問題永遠不會直接提交給模型。首先，系統會從你的文檔中提取正確的事實，然後將這些事實與問題一併提交給模型，最後在任何人看到答案之前，檢查草稿答案是否與這些事實一致。'紮根'是基礎，而最強大的系統還會增加兩個額外層級。微軟Azure的'紮根性檢測'功能能夠識別答案中任何未被來源支持的部分，並在用戶看到前進行重寫。NVIDIA開源的NeMo Guardrails為RAG系統增加了事實核查機制。第二層是所有保障中最可靠的一環：對於高風險決策，必須引入人工參與閉環。

企業準確性操作手冊中應包含哪些內容?

- **確保每個答案都有依據。** 通過RAG將AI連接到可信且最新的數據，使其基於事實而非記憶進行回答。
- **展示信息來源並添加護欄。** 要求AI明確引用每個答案的來源，並通過自動化檢查來識別或修正缺乏支持的陳述。
- **保持人工參與。** 對涉及金錢、法律、醫療或面向客戶決策的情況，必須由人工在發佈前進行審核。
- **從數據入手。** 清潔、標註良好且適合AI處理的數據是上述所有環節得以運行的輸入基礎。

Lifewood是如何將數據驅動的準確性變為現實的?

可信的AI建立在可信的數據之上，而這正是Lifewood所專注的領域。自2004年成立以來，公司已轉型為專注於AI數據服務的專業機構，目前業務覆蓋全球30多個國家和40多個交付中心。它結合了全球範圍的人力資源與工業化的方法論，以及其專有的LiFT平臺。

AIGC 指人工智能生成內容：生成模型所生產出的文本、圖像、音頻和視頻。Lifewood 在兩端均支持 AIGC，既準備高質量數據以確保生成內容的準確性，又實施全天候的人工參與閉環質量控制，確保輸出內容的可靠性。

- **數據採集。** 覆蓋文本、音頻、圖像和視頻的多語言、多模態數據採集，支持50多種語言。
- **標註與標籤化。** 標註、打標籤、轉錄和情感分析：模型學習的結構化真實基準。
- **大語言模型訓練數據。** 監督微調數據集、人工偏好（RLHF）數據以及模型評估數據集。
- **AIGC與問答服務。** 提供企業級AI生成內容，包括大規模視頻內容，背後均配有全職的人工審核與驗證。

幻覺問題通常源於數據質量，而非僅是模型問題。Lifewood的專業團隊從源頭入手解決該問題：對檢索到的文檔進行結構化和驗證，解決57%數據未達到AI可用性的組織問題；構建行業和語言特定的數據集以及評估集，以教會模型準確回答並承認不確定性；並提供覆蓋50多種語言的全球人力作為驗證層，彌補自動化檢查所遺漏的錯誤。

管道從數據採集開始，經過標註、訓練與評估，再到人工質量審核。這些階段共同構建了一個有據可依的知識庫，支持助手基於已驗證的事實進行回答。當AI出現偏差時，錯誤會反向流回數據流程，從而使下一個版本更加準確。

企業級AI準確性的下一步是什麼?

守護智能體：用AI檢查AI。 一個快速興起的想法是，AI監督其他AI。守護代理會審查、監控，並能夠阻止另一個模型產生危險的輸出。

可直接用於AI的數據成為優先事項。 關注點正從炫目的模型轉向基礎層面：乾淨、受良好治理的數據。鑑於57%的組織表示其數據不具備AI可用性，那些在此領域投入的組織將悄然領先。

最終結論是什麼?

幻覺並非表明AI出了問題，而是語言模型工作方式中可預測且可控的特徵。這一策略被麥肯錫、高德納、OpenAI、谷歌、AWS、IBM、NVIDIA和微軟等機構廣泛認同。將模型建立在可信數據之上，展示信息來源，設置防護機制，保持人工參與閉環，並對整個流程進行治理。

每一步背後都存在一個共同要求：高質量、充分準備、具備AI可用性的數據。由於57%的組織缺乏此類數據，且51%的組織已報告出現負面的AI影響，工作必須從這裡開始。突破點並非更聰明的模型，而是更優質的數據，且需以謹慎方式處理。你無法通過提示工程來解決數據問題——準確性必須在數據層面建立，遠在模型開口之前就已形成。

資料來源

- 《福布斯》／加拿大不列顛哥倫比亞省民事爭議解決審裁處，*Moffatt 訴加拿大航空公司案*，2024 BCCRT 149。
- IBM. *什麼是檢索增強生成（RAG）？*
- OpenAI. *語言模型為何會產生幻覺。* 2025.
- McKinsey & Company. *人工智能現狀*, 2025年全球調查。
- 普華永道與墨爾本大學. *信任、態度與人工智能的使用：一項全球研究 2025。*
- 哈佛商業評論. *你的公司有哪些人工智能噩夢？* 2026.
- Dahl, M. et al. *大型法律虛構：大型語言模型中法律幻覺的剖析。* 法律分析期刊 16, 64 (2024).
- Magesh, V. et al. *無幻覺？評估領先的人工智能法律研究工具的可靠性。* 經驗法律研究期刊 22, 216–242 (2025).

- 谷歌雲、亞馬遜AWS、英偉達、微軟Azure人工智能內容安全，以及GitHub / 英偉達NeMo防護機制產品文檔。
- Gartner. *人工智能 hype 週期*, 2025。

編輯說明：本文解釋了廣泛使用的技術如何減少AI錯誤。沒有任何方法能完全消除這些錯誤，結果取決於每個組織的數據和具體應用場景。上述資料中的統計數據均來自公開可獲取的來源，並在文中註明。