

>_

統計學家George E. P. Box: All models are wrong, but some are useful.
此句的意義對於現今的LLM發展有何啟發性的意義, 特別是大模型幻覺,
對現實世界的理解方面, 請深入探討.

ChatGPT >_

你希望我從哪個角度深入探討這句話對大語言模型(LLM)發展的啟發? 例如:

1. 技術層面: 如模型架構、訓練數據與幻覺現象的關聯。
2. 哲學與認知層面: 如對知識表示、真實性與語言生成的理解。
3. 實務應用層面: 如在醫療、法律、教育等領域如何看待模型的不準確性。

也請告訴我你希望分析的深度與格式, 例如是一篇評論文章、演講稿草稿, 還是摘要式討論。

>_

技術層面: 如模型架構、訓練數據與幻覺現象的關聯。

ChatGPT >_

好的, 我將從技術層面深入探討George E. P. Box的名言「All models are wrong, but some are useful」對當代大語言模型(LLM)發展的啟發, 特別聚焦於模型架構、訓練數據與幻覺現象之間的關聯。

我會整理出與此主題相關的最新研究與專家觀點, 稍後回覆你詳細的分析內容。

Research completed in 16m·20 個資料來源· 64 次搜尋

大語言模型幻覺與模型限制的技術啟示

「所有模型都是錯的, 但有些是有用的」對大型語言模型的啟示

引言: 模型的錯誤性與實用性

統計學家George E. P. Box的名言「All models are wrong, but some are useful」強調了模型本質上的近似性: 任何模型都無法完全呈現現實, 但仍能提供實用價值zh.wikipedia.org。

這句話在當代人工智慧、特別是大型語言模型(LLM)的發展中依然適用。LLM如GPT-4

等在語言理解與生成上取得突破，但終究只是對真實世界的簡化模擬zh.wikipedia.org。以下我們從技術層面探討該名言對LLM的啟示，包括模型架構設計的內在限制、訓練數據偏誤引發的幻覺問題，以及如何在生成內容時平衡模型的「有用」與「錯誤」，最後介紹當代緩解這些問題的技術策略（如RAG、微調、對齊）及其成效與侷限。

模型架構的設計限制：理解與表徵的瓶頸

大型語言模型的架構（例如Transformer架構）雖然強大，但其設計使其對世界的理解本質上仍有局限。LLM並非真正「理解」語義或具有常識推理能力，而主要是透過大量語料中學習到的統計模式來預測詞語ve3.global。換言之，模型生成答案更多是根據字詞在訓練數據中出現的機率，而非基於對問題含義的深刻領會ve3.global。有研究將此形容為一種高級自動完成：模型的目標是產生在形式上合理的內容，並不會主動驗證內容的真實性mitsloanedtech.mit.edu。因此，LLM給出的回應往往聽起來合理但未必正確，在需要深度理解或邏輯推理的情境下尤為如此ve3.global。另外，當前LLM的上下文記憶能力也受架構限制：模型一次處理的文本長度有限，無法將無限制的背景知識納入考量

montecarlodata.com。模型的知識基礎也是相對靜態的——訓練完成後參數中的知識不會隨時間自動更新ve3.global。這意味著模型對訓練語料中未涵蓋的新知或變化，缺乏及時的反映能力。因此，在架構上LLM始終是對真實世界的簡化映射，正如Box所說「地圖並非領土」，它有解析度和範圍的限制。

訓練數據組成與偏誤：幻覺現象的成因

大型語言模型的幻覺（hallucination）現象，指模型自信輸出與事實不符的內容，與其訓練數據的品質和偏誤有直接關聯。模型從網際網路等大量語料中學習語言，但這些語料中摻雜了錯誤資訊和偏見。模型缺乏人類辨別真偽的能力，如果訓練資料中存在謬誤或偏頗觀點，模型往往也會模仿出類似的內容mitsloanedtech.mit.edu。正如MIT的一份分析所指出：「生成式AI之所以產生虛假和偏見內容，源於訓練數據本身良莠不齊、模型側重模式匹配的設計，以及AI技術內在的局限」mitsloanedtech.mit.edu。其中，訓練資料的偏差是一個顯著因素：如果模型的訓練集帶有系統性的偏見或不完整，模型就可能憑藉這些片面的經驗「想像」出不存在的模式或細節，使偏見在輸出中放大ibm.com。此外，模型在生成內容時並沒有連結真實世界的機制，即便完全基於訓練資料，它也可能以全新組合方式產生前所

未見的錯誤陳述mitsloanedtech.mit.edu。這種情況下，模型輸出雖語氣自信但實則內容虛構，令人誤以為是有依據的資訊。現實案例亦證實了這點：例如2023年美國 *Mata v. Avianca* 訴訟中，一位律師使用ChatGPT查找判例，結果ChatGPT杜撰了數個並不存在的法院案例和引文，甚至信誓旦旦地稱可在法律資料庫查到這些案例mitsloanedtech.mit.edu。可見，當訓練語料缺乏某領域的正確資訊時，模型寧可編造一個答案（哪怕錯誤）也不願承認「不知道」，這正是幻覺現象的重要成因。

「有用性」與「錯誤性」的平衡：生成內容的取捨

在應用大型語言模型時，我們需要在模型輸出的實用價值和其可能的錯誤之間取得平衡。“有用”通常意味著模型願意提供詳盡的回答，而“不那麼錯”則意味著模型在不確定時應保持謹慎。極端地追求零錯誤會導致模型過於保守，頻繁地拒絕回答或只給出非常模糊的回應，削弱了實用性；反之，一味追求豐富輸出又可能增加錯誤內容的風險。

OpenAI的一項比較研究揭示了這種權衡：一個較早的模型為了提高問答的表面準確率，在不確定時選擇猜測答案，結果雖然命中了一些答案但同時大幅提升了出錯率

openai.com。相反，經過調整的新模型傾向於在沒把握時選擇不作答或表達不確定，錯誤率明顯降低但回答率也隨之下降openai.com。因此，模型積極回答與避免謬誤之間存在此消彼長的關係。我們可以根據應用場景來調節這個度：在創意寫作等場景下，一定程度的「幻覺」未嘗不是模型展現想像力的一種方式，有觀點甚至認為這類幻覺提供了新奇的思路和靈感來源——模型產生的某些錯誤其實是對現實的另類建模，蘊含著啟發性

marginalrevolution.com。相對地，在醫療、法律等嚴謹領域，對錯誤零容忍是必要的，此時模型應優先確保可靠性，即便因此略顯保守。技術上也可以透過調節參數來影響這種平衡，例如降低生成隨機性（temperature）能讓模型輸出更一致且貼近事實的內容，而提高temperature則增強輸出的發散性與創造力mitsloanedtech.mit.edu。實驗表明，將temperature設置在0~0.3區間時，模型回答會更專注嚴謹；而temperature調高到0.7以上時，模型更可能天馬行空地「發揮」，適合腦力激盪但伴隨更高出錯風險mitsloanedtech.mit.edu。總之，根據用途調整模型的「想像力」大小，有助於在有用性與正確性之間取得適當的平衡。

緩解錯誤與提升實用性的技術策略

面對上述挑戰，研究者和工程師提出了多種技術方案來降低LLM的幻覺率、提高事實正確性，同時盡量保持模型的有用性。以下介紹幾種重要的策略及其效果與限制：

檢索增強生成(RAG)

策略描述：RAG(Retrieval-Augmented Generation)在模型生成回答前先執行資訊檢索，給模型提供與問題相關的外部資料作為輔助。這彌補了LLM參數中知識的不足，將動態知識引入模型的上下文。效果：RAG有效降低了模型產生陳舊或虛假資訊的概率。因為當模型遇到知識盲區時，會優先搜尋真實資料而不是憑空編造細節，從而大幅減少胡亂猜測導致的錯誤^{montecarlodata.com}。具體而言，透過檢索最新的資料，模型能提供即時且有依據的答案。例如，詢問一個傳統模型昨天的股價走勢，它只能依賴舊知識猜測，而結合RAG的模型則會檢索最新的財經數據後再回答^{montecarlodata.commontecarlodata.com}。研究表明，這種將檢索與生成結合的架構，顯著提升了回答的事實準確性，也增進了使用者對AI答案的信任^{mitsloanedtech.mit.edu}。同時，由於答案引用了外部來源，RAG有助於提供溯源依據：使用者可以追查模型引用了哪些文件來形成答案，提高了透明度和可驗證性

^{montecarlodata.com}。限制：雖然RAG強化了模型的知識，但也引入了新的複雜性。首先，實現RAG需要打造完整的檢索基礎設施，包括維護文件資料庫、向量索引以及高效的語義搜索機制等，這對工程團隊提出了較高的技術門檻和維運成本^{montecarlodata.com}。其次，模型的上下文窗口長度限制依然存在：檢索到的相關內容都必須塞進提示中供模型參考，如果問題涉及資訊量很大或多個長文檔，系統必須對資料進行切分和過濾，只選取最相關的片段給模型^{montecarlodata.com}。不當的切片策略可能導致關鍵資訊遺漏或上下文不連貫。最後，如果知識庫中沒有包含相關答案，模型仍可能無法給出正確解答(此時理想情況下應該拒答或表示不確定)。因此RAG的效能取決於知識庫的完備性：它在資料範圍內表現優異，但對於超出知識庫範圍的問題仍然愛莫能助。此外，RAG在整合檢索結果時也可能引發一致性問題——模型需要將多來源資訊與提問整合，在這過程中若資料相互矛盾或品質不佳，仍可能導致錯誤結論。

微調模型(Fine-tuning)

策略描述：微調是指在預訓練模型的基礎上，使用特定領域或任務的資料對模型進行再訓練調優，使其更貼合特定需求。這包含兩種形式：(1)領域微調：例如以醫學文本微調LLM，讓模型熟悉醫學知識和行文風格；(2)對齊微調：以人類偏好反饋資料(如安全守則或高質量對話)來訓練模型，這部分在下一節討論。效果：微調能顯著提升模型在特定領域的表現。透過調整內部權重，模型對該領域的專業知識和語用規則有了更強的表徵

能力montecarlodata.com。例如，微調後的法律助理模型不僅瞭解法律術語，還學會了何時使用「據此」「茲此」等句式，如何組織法律論點，甚至能熟練引用常見的判例montecarlodata.com。同樣地，微調過的醫學模型會以專業醫學縮寫和正式語氣給出診斷建議，而不會像通用模型那樣泛泛而談montecarlodata.com。這種專業化使模型在該領域的回答更準確可靠，同時在格式和語氣上更符合預期。此外，微調還能提高效率：針對特定任務微調後的中小型模型，有時可媲美體量大得多的通用模型表現，達到以小模型實現高性能的目的

montecarlodata.com。限制：微調也有明顯局限和代價。首先是數據與算力需求高：有效的微調需要大量經過精心整理和標註的訓練數據，以及昂貴的計算資源montecarlodata.com。準備領域資料集往往耗費數週甚至數月的人力，而對上億參數模型進行微調需要強大的GPU/TPU支援，一次完整訓練可能花費數萬美元的計算成本

montecarlodata.com。這使得非大型機構難以頻繁微調模型。其次，過度專精風險：模型微調後雖在特定領域表現出色，但可能過擬合於該領域語料montecarlodata.com。一個在法律文本上深度微調的模型，也許閱讀合同得心應手，但可能無法再準確回答一般常識性的問題，甚至連先前一些基本能力也退化了montecarlodata.com。模型可能記住訓練集中出現的片段，對略有變化的輸入反而表現不佳。再次，知識時效性問題：如果目標領域的知識更新很快，微調模型很難時時保持最新。當新的醫學研究或法律條文出現後，除非獲得新資料重新訓練，否則模型仍沿用舊知識montecarlodata.com。而每次更新微調都意味著重新蒐集資料並跑一次昂貴的訓練流程，週期長成本高。在快速變動的領域中，模型很可能還沒來得及更新就已過時montecarlodata.com。相比之下，RAG只需將新文件加入資料庫即可立即讓模型檢索使用，維護成本更低。最後，透明度和信任問題：微調後模型給出的答案往往無法附帶來源，因為資訊都隱含在參數中而非顯式文件montecarlodata.com。這對高風險應用是個隱患——例如醫生從模型那裡得到診斷意見，卻無從確認這個結論源於哪篇醫學研究。缺乏可追溯性使得使用者難以對模型產生完全信任montecarlodata.com。綜上，微調在提升模型專業性的同時，也引入了模型泛化能力降低、知識更新不便等問題，需要根據應用情境權衡使用。

模型對齊與人類反饋 (Alignment)

策略描述：模型對齊是指調整AI行為使其符合人類期望和價值觀的一系列方法。其中最典型的是以人類反饋進行強化學習 (RLHF)：蒐集人類偏好 (例如對回答是否滿意、是否

有害的不同行為打分)，據此對模型進行訓練調優，使其傾向於產生讓人類評價更高的回應。此外，還有如Anthropic提出的***“憲法式AI”（讓模型遵循一組人道原則自我調整）等對齊方法。本節主要關注RLHF等與減少幻覺、提高有用性相關的技術。效果：經過對齊優化的模型通常在安全性、禮貌程度以及遵循指令方面有明顯改善，同時也嘗試變得更誠實和準確**。

例如，OpenAI透過RLHF訓練的InstructGPT和ChatGPT系列模型，比未對齊的GPT-3更不容易產生攻擊性內容，並在真實性評估（如TruthfulQA基準）上有一定提升。對於幻覺問題，對齊策略的一個重點是在模型不確定答案時鼓勵誠實表達無知，而非編造。openai.com就闡明了這種原則：模型若無把握，最好直言「我不確定」或要求更多資訊，比自信地給出可能錯誤的回答要來得負責openai.com。實踐中，我們已看到對齊帶來的一些改善：最新一代經過強化學習對齊的模型（如GPT-4等）相較早期模型，幻覺明顯減少，尤其在推理題上表現更穩健openai.com。OpenAI的團隊指出，即將推出的更先進模型（暫稱GPT-5）在對齊和推理方面進一步優化，顯著降低了幻覺率openai.com。這說明適當的對齊訓練確實能讓模型更謹慎地面對自身知識盲區，提升整體可靠性。限制：然而，對齊並非靈丹妙藥，仍存在多方面局限。首先，幻覺問題依然存在：再多的訓練也無法讓一個有限的模型成為完美知識的源頭。OpenAI坦承，幻覺仍是所有大型語言模型尚未完全解決的基本挑戰

openai.com。模型或許會更少犯錯，但在面對未見過的知識或模稜兩可的問題時，難免仍會產生不正確的內容。我們對此需要保持警惕，不應因模型表現提升就掉以輕心。其次，RLHF這類對齊方法本身可能引入新的問題，例如過度自信或迎合偏差。有研究發現，經RLHF微調後的模型在表達置信度時的校準變差了——也就是說，模型可能對自己的回答更有信心，即便那些回答實際上是錯的lilianweng.github.io。這或許是因為在人類反饋過程中，評價者更喜歡聽到確定的回答而非「我不確定」，導致模型學到一種迎合人類偏好的表達方式，犧牲了對不確定性的誠實呈現lilianweng.github.io。另外，對齊過程若處理不當，還可能將人類的主觀偏見引入模型決策，使模型在某些議題上呈現趨附人意的傾向而非中立求真。最後，對齊訓練往往與模型其它目標存在拉鋸：一味壓制模型的不當輸出，可能同時抑制了模型的創造力或有用性。例如，有些對齊嚴格的模型會過度避免回答任何具有一點風險的問題，導致使用體驗受影響。總的來說，模型對齊在提升LLM實用性的同時，也是在模型的「錯誤性」上進行權衡取舍。透過不斷調整對齊策略，我們希望找到既不失實用價值又將錯誤降至可接受範圍的平衡點。結語：實用至上，持續改進

George Box的箴言提醒我們，模型終究是對現實的近似，「錯誤」在所難免，但關鍵在於其實用性是否足以產生價值zh.wikipedia.org。對於大型語言模型，我們應抱持同樣的心態：既不要以絕對完美的標準苛求它們，也不能忽視其內在局限帶來的風險。在技術上，我們通過改進模型架構、嚴謹選取和整理訓練數據，以及採用檢索增強、微調、對齊等方法，不斷縮小模型與真實世界的差距。同時，我們也需要建立配套的監督與評估機制，在模型出錯時及時發現並糾正。正如Box所強調的，模型的價值最終取決於它帶給我們的實用效益，而非它是否完美無缺zh.wikipedia.org。大型語言模型的開發者正是秉持這一理念，在接受「所有模型都是錯的」的前提下，最大限度發揮「但有些是有用的」— 致力於讓LLM成為對人類有益且值得信賴的工具。

參考來源：

- Box, G. E. P. 等人在統計模型領域的觀點
- zh.wikipedia.org
- zh.wikipedia.org
-
- VE3 Global 技術團隊對LLM世界模型潛力與限制的分析
- ve3.global
- ve3.global
-
- MIT教育科技團隊對生成式AI偏誤與幻覺成因的報告
- mitsloanedtech.mit.edu
- mitsloanedtech.mit.edu
-
- IBM對AI幻覺現象及其與訓練資料偏誤關係的解讀
- ibm.com
-
- OpenAI關於LLM幻覺問題的研究與準則(2025)
- openai.com
- openai.com
- openai.com
-
- Tyler Cowen 對ChatGPT幻覺創造力的評論(2023)
- marginalrevolution.com
-
- Lilian Weng 關於減少幻覺的方法與模型校準的觀察(2024)
- lilianweng.github.io
-
- Monte Carlo Data 關於RAG與微調取舍的技術博客(2024)
- montecarlodata.com
- montecarlodata.com
- montecarlodata.com
-
- 專家對微調效益與侷限的分析報告
- montecarlodata.com
- montecarlodata.com

- montecarlodata.com
- montecarlodata.com
-
- MIT Sloan對利用RAG提升事實性的建議
- mitsloanedtech.mit.edu
-
- 真實案例: *Mata v. Avianca* 案中LLM幻覺的教訓
- mitsloanedtech.mit.edu
-

資料來源

引用

zh.wikipedia.org

[所有模型都是錯的 - 维基百科, 自由的百科全书](#)

[维基百科, 自由的百科全书](#)

ve3.global

[LLMs and World Modelling: Exploring the Potential and Limitations](#)

[Lack of True Understanding](#)

mitsloanedtech.mit.edu

[When AI Gets It Wrong: Addressing AI Hallucinations and Bias - MIT Sloan Teaching & Learning Technologies](#)

[reproduce any falsehoods or biases present in that data \(Weise & Metz, 2023\). 2. Limitations of Generative Models: Generative AI models function like advanced autocomplete tools: They're designed to predict the next word or sequence based on observed patterns. Their goal is to generate plausible content, not to verify its truth. That means any accuracy in their outputs is often coincidental. As a result, they might produce content that sounds reasonable but is inaccurate \(O'Brien, 2023\). 3.](#)

[Inherent Challenges in AI Design: The technology behind generative AI tools](#)

montecarlodata.com

[RAG Vs. Fine Tuning: Which One Should You Choose?](#)

[# Context window limitations](#)

ve3.global

[LLMs and World Modelling: Exploring the Potential and Limitations](#)

[Static Knowledge Base](#)

mitsloanedtech.mit.edu

[When AI Gets It Wrong: Addressing AI Hallucinations and Bias - MIT Sloan Teaching & Learning Technologies](#)

[1. Training Data Sources: Generative AI models are trained on vast amounts of internet data. This data, while rich in information, contains both accurate and inaccurate content, as well as societal and cultural biases. Since these models mimic patterns in their training data without discerning truth, they can reproduce any falsehoods or biases present in that data \(Weise & Metz, 2023\). 2. Limitations of Generative Models: Generative AI models function like advanced autocomplete tools: They're designed to predict the next word or sequence based on observed patterns. Their goal is to generate plausible content, not to verify its truth. That means any accuracy in their outputs is](#)

mitsloanedtech.mit.edu

[When AI Gets It Wrong: Addressing AI Hallucinations and Bias - MIT Sloan Teaching & Learning Technologies](#)

In short, the “hallucinations” and biases in generative AI outputs result from the nature of their training data, the tools’ design focus on pattern-based content generation, and the inherent limitations of AI technology. Acknowledging and addressing these challenges will be essential as generative AI systems become more integrated into decision-making processes across various sectors.

[ibm.com](#)

[What Are AI Hallucinations? | IBM](#)

hallucination in machine learning algorithms is input bias. If an AI model is trained on a dataset comprising biased or unrepresentative data, it may hallucinate patterns or features that reflect these biases.

[mitsloanedtech.mit.edu](#)

[When AI Gets It Wrong: Addressing AI Hallucinations and Bias - MIT Sloan Teaching & Learning Technologies](#)

3. Inherent Challenges in AI Design: The technology behind generative AI tools isn’t designed to differentiate between what’s true and what’s not true. Even if generative AI models were trained solely on accurate data, their generative nature would mean they could still produce new, potentially inaccurate content by combining patterns in unexpected ways (Weise & Metz, 2023).

[mitsloanedtech.mit.edu](#)

[When AI Gets It Wrong: Addressing AI Hallucinations and Bias - MIT Sloan Teaching & Learning Technologies](#)

For an example of how AI hallucinations can play out in the real world, consider the legal case of Mata v. Avianca. In this case, a New York attorney representing a client’s injury claim relied on ChatGPT to conduct his legal research. The federal judge overseeing the suit noted that the opinion contained internal citations and quotes that were nonexistent. Not only did the chatbot make them up, it even stipulated they were available in major legal databases (Weiser, 2023).

[openai.com](#)

[Why language models hallucinate | OpenAI](#)

100%

[marginalrevolution.com](#)

[What is an optimum degree of LLM hallucination? - Marginal REVOLUTION](#)

The hallucinations also give me ideas and show me alternative pathways. “What if...?” They are a form of creativity. Many of these hallucinations are simple factual errors, but many others have embedded in them alternative models of the world. Interesting models of the world. Ideas and inspirations. I feel I know what question to ask or which task to initiate.

[mitsloanedtech.mit.edu](#)

[When AI Gets It Wrong: Addressing AI Hallucinations and Bias - MIT Sloan Teaching & Learning Technologies](#)

transparency and accuracy in complex tasks (Wei et al., 2022). 5. Adjust the Tool’s Temperature: Temperature is a setting that controls how random or creative the model’s responses are. In tools that allow you to adjust it, using a low temperature (e.g., 0–0.3) produces more focused, consistent, and factual outputs—especially for well-defined prompts. A higher temperature (e.g., 0.7–1.0) encourages more varied and imaginative responses, making it better suited for open-ended tasks like brainstorming or storytelling.

montecarlodata.com

[RAG Vs. Fine Tuning: Which One Should You Choose?](#)

[# Reduced hallucinations](#)

montecarlodata.com

[RAG Vs. Fine Tuning: Which One Should You Choose?](#)

[information or access new data, similar to someone trying to answer current events questions using only an old encyclopedia. This limitation led Meta AI researchers to introduce RAG in 2020 as a solution.](#)

montecarlodata.com

[RAG Vs. Fine Tuning: Which One Should You Choose?](#)

[The practical implications are significant. If you ask a traditional language model about yesterday's stock prices or a company's latest earnings report, it can only guess based on outdated patterns in its training data. A RAG-powered model, however, would search its connected database for the most recent financial information and deliver an accurate, timely response. This approach transforms AI applications from static knowledge repositories into dynamic tools that can access and incorporate fresh information whenever needed.](#)

mitsloanedtech.mit.edu

[When AI Gets It Wrong: Addressing AI Hallucinations and Bias - MIT Sloan Teaching & Learning Technologies](#)

[3. Use Retrieval-Based Tools: Some generative AI tools are built with Retrieval-Augmented Generation \(RAG\) architectures. This means they'll retrieve relevant information from trusted sources—such as your syllabus, research article, or case PDF—before generating output. Research has shown that RAG improves both factual accuracy and user trust in AI-generated answers \(Li et al., 2024\).](#)
[4. Use Clear and Structured Prompts: The quality of AI output is closely tied to how specific your input is. Vague prompts often lead to vague—or even inaccurate—answers. You can reduce this risk by setting clear expectations and](#)

montecarlodata.com

[RAG Vs. Fine Tuning: Which One Should You Choose?](#)

[# Traceability and verification](#)

montecarlodata.com

[RAG Vs. Fine Tuning: Which One Should You Choose?](#)

[# Building and maintaining retrieval infrastructure](#)

montecarlodata.com

[RAG Vs. Fine Tuning: Which One Should You Choose?](#)

[Because the model's weights are adjusted during fine-tuning, the resulting model can generate responses that are more accurate and contextually appropriate for the domain it's tuned on. A fine-tuned legal AI doesn't just know legal terms; it understands when to use "whereas" clauses, how to structure legal arguments, and which precedents typically apply to different situations. Similarly, a medical AI fine-tuned on clinical notes learns not just medical vocabulary but also the concise, specific way doctors document patient encounters. The model tends to use the correct jargon naturally, follow desired formats without explicit prompting, and handle niche questions that would stump a generic model.](#)

montecarlodata.com

[RAG Vs. Fine Tuning: Which One Should You Choose?](#)

Efficiency gains through model optimization

montecarlodata.com

[RAG Vs. Fine Tuning: Which One Should You Choose?](#)

Data and computational requirements

montecarlodata.com

[RAG Vs. Fine Tuning: Which One Should You Choose?](#)

Fine-tuning an LLM requires a lot of high-quality, labeled data and computational resources that many organizations struggle to provide. Preparing the dataset can be labor-intensive, involving extensive data collection, cleaning, and labeling processes that might take weeks or months. A healthcare organization fine-tuning a model for medical diagnosis assistance might need thousands of accurately labeled case studies, each verified by medical professionals. The training process itself demands powerful GPUs or TPUs, especially for large models, with costs that can reach tens of thousands of dollars for a single fine-tuning run. Even with parameter-efficient methods,

montecarlodata.com

[RAG Vs. Fine Tuning: Which One Should You Choose?](#)

Risk of overfitting

montecarlodata.com

[RAG Vs. Fine Tuning: Which One Should You Choose?](#)

Maintenance and knowledge updates

montecarlodata.com

[RAG Vs. Fine Tuning: Which One Should You Choose?](#)

Maintaining a fine-tuned model proves challenging when domain knowledge evolves. If new medical research emerges or laws change, you must update the training data and re-train the model to keep it current. This retraining cycle can take days or weeks and cost thousands of dollars in compute resources, compared to RAG's approach of simply adding new documents to a database. Organizations in rapidly changing fields might find themselves in a constant cycle of retraining, with models becoming outdated before the next version is ready. The lag between identifying needed updates and deploying a retrained model can leave users

montecarlodata.com

[RAG Vs. Fine Tuning: Which One Should You Choose?](#)

Lack of source attribution

openai.com

[Why language models hallucinate | OpenAI](#)

For questions where there is a single "right answer," one can consider three categories of responses: accurate responses, errors, and abstentions where the model does not hazard a guess. Abstaining is part of humility, one of OpenAI's core values. Most scoreboards prioritize and rank models based on accuracy, but errors are worse than abstentions. Our Model Spec ([opens in a new window](#)) states that it is better to indicate uncertainty or ask for clarification than provide confident information that may be incorrect.

openai.com

[Why language models hallucinate | OpenAI](#)

ChatGPT also hallucinates. GPT-5 has significantly fewer hallucinations especially when reasoning, but they still occur. Hallucinations remain a fundamental challenge for all large language models, but we are working hard to further reduce them.

lilianweng.github.io

[Extrinsic Hallucinations in LLMs | Lil'Log](#)

[the model is expected to demonstrate the right level of confidence.](#)

zh.wikipedia.org

[所有模型都是錯的 - 维基百科，自由的百科全书](#)

[\[编辑\]](#)

所有資料來源

[openai](#)

[2](#)

[arxiv](#)

arxiv.labs.arxiv

[2](#)

lilianweng.github

[3](#)

[ve3](#)

[2](#)

[amazon](#)

[montecarlodata](#)

[3](#)

en.wikipedia

zh.wikipedia

[2](#)

[marginalrevolution](#)

[ibm](#)

mitsloanedtech.mit

ChatGPT 可能會出錯。請查核重要資訊。