

DeepSeek下載火爆 一周超過谷歌Gemini兩年

大公報專訪AI權威 破譯「深度求索」成功密碼



DeepSeek風暴7

近期，國產AI大模型DeepSeek橫空出世，僅以幾十分之一的成本，就達到與全球領先的ChatGPT最新o1版本推理模型相當的實力。目前，DeepSeek下載量持續攀升，據研究公司SimilarWeb估計，在短短一周內，DeepSeek網站的用戶數量就超過了谷歌的Gemini聊天AI，而後者已經存在了近兩年。一時間，各界眾說紛紜：DeepSeek為什麼這麼牛？何以如此？是創新還是蒸餾？……諸如此類的疑問頻出。

人工智能專家高飛長期跟蹤研究ChatGPT與DeepSeek的發展軌跡，他在接受大公報獨家專訪時，從技術肌理與論文研究原點出發，破譯「深度求索」成功的基因密碼，揭密DeepSeek低成本、高性能的原因。他還表示，DeepSeek-R1「出圈」，得益於高性能、低成本、強開放三個因素，這與美國閉源、收費、高成本的ChatGPT形成鮮明對比。

大公報記者 蘇雨潤、凱雷北京報道

三大創新凝煉DeepSeek 「東方神秘力量」

大公報：DeepSeek是如何實現低成本的？

高飛：DeepSeek的低成本得益於其創新性，它打破了「只有靠最先進硬件才能訓練出前沿模型」的傳統觀念。DeepSeek團隊利用有限的資源，通過創新的方式，訓練出了以「測試時間計算」方式運行的推理模型，並對標OpenAI最先進的o系列模型，實現了看似「不可能」的「低成本+高性能」極限組合，並提供開源代碼，讓企業、個人可以在本地算力設施上部署免費使用。所以，它的低成本不是「因」，而是「果」，且是「果」之一。

大模型的「大」，其實是大算力、大參數、大數據。每一個大，都意味着高成本。DeepSeek在約束條件下，做了極限創新，減少算力需求、減少參數數量、降低數據規模。這是在資源上做減法，但卻不對性能有任何妥協。所以，DeepSeek不僅「低成本」，還有「高性能」，以及開源模式的「強開放」。

這次OpenAI跟隨DeepSeek-R1發布o3 Mini模型，CEO山姆·奧特曼承認其對於開源的判斷有誤，站在了「歷史錯誤的一邊」。言外之意，OpenAI可能也會在未來考慮對其模型全部或者部分開源。因此，DeepSeek-R1模型的成功，不僅是技術創新的成功，更是技術開放的成功。

實現強化學習法 讓用家看到「內心戲」

大公報：DeepSeek-R1兼具低成本和超高性能，更在低成本的前提下實現對標OpenAI o系列模型，這究竟是怎麼做到的呢？

高飛：DeepSeek-R1模型代表了高性能、新思維的突破。大家知道，OpenAI的o系列模型是一種測試時間計算（Test-Time Computation）模型，模型在推理階段（Inference Phase）會執行計算的方式。表現在使用上，就是模型不會立刻回答用戶的問題，而是根據問題難度，經過一番思考（通常是幾十秒），給出高質量答案。

DeepSeek R1使用了一種創新的強化學習（RL）方法，實現了該種模型的訓練。最終實現的效果是，當你提問DeepSeek-R1，它看似是最終給你一個答案，但其實模型有大量的「內心戲」，是經過反覆推敲、多步思考，最終給出一個相對完美的答案。在學術上，這個過程叫思維鏈（Chain of thought）。

大公報：可否從技術肌理層面為我們解讀，DeepSeek的高性能主要源於何種技術創新呢？

專家剖析DeepSeek技術

謬誤1

DeepSeek是在ChatGPT大模型上通過「蒸餾」產生的，這涉及「偷取」知識產權，且任何人都可以以低成本蒸餾出「精華」。

高飛：否。模型蒸餾並不是一項新技術，目前沒有任何公司利用蒸餾方法，訓練出超過其他模型性能的產品。這兩年，擁有高端芯片、強大算力、在ChatGPT上蒸餾數據的模型，不下幾十個，沒有一個能「跑」出類似的效果，都達不到DeepSeekR1強大的性能。而且，DeepSeek的技術秘密是公開的，它既模型開源，又在原始技術論文中公開了細節。

減少消耗 分工明確「餐廳模式」

高飛：DeepSeek所擁有的「東方神秘力量」——高性能，主要來自混合專家模型（MoE）、多標記預測（MTP）和多頭潛在注意力機制（MLA）等三大技術的創新。其中，MoE技術實現了DeepSeek在參數上的精簡。例如，去年底發布的DeepSeek-V3，就是DeepSeek團隊採取混合專家模型的模式，將大模型分成多個「專家」子模型，將DeepSeek-V3的671B參數進行拆解，每個子模型的參數量大約只有37B（總參數的1/20），且擅長不同領域的知識。這樣，針對不同的輸入數據，模型會動態地選擇最合適的部分專家來參與計算，自然就減少了資源消耗。

大公報：可以舉一個形象的例子，解釋一下MoE技術是如何運行的嗎？

高飛：訓練大模型就像點菜，美國的主流大模型就像一個有100個檔口的大牌檔，用戶點一份披薩，所有廚師、服務生都要動起來。而MoE技術就相當於組織分類，用戶同樣點一份披薩，西餐檔口只調動相應的西餐廚師和服務生，其他檔口並不受擾動，因此忙而不亂，井然有序。從粗放式管理到精細化模式，即專家模式，其實就是提升模型的組織創新力。

同步炮製 後廚「預判式做飯」

高飛：如果說，MoE技術讓模型盡可能用更少的參數工作，是空間上的優化，那麼MTP技術就是時間上的優化，它讓模型用同樣的資源做更長期的工作。MTP技術可以讓模型不是每次預測生成一個Token（文本拆分的最小單元），而可以在每個位置預測多個未來Token，這就更大程度利用了計算資源，增加了訓練信號密度，提高了訓練效率。就像讓餐廳檔口做第一道菜時，就同步準備第二、三道菜，用一份算力做更多工作，將資源利用到極致。

化繁為簡「例湯白飯式」歸類

高飛：多頭潛在注意力機制（MLA）技術，就是將原始高維特徵壓縮到一個較低維度的潛在空間（潛在向量），再通過上投影矩陣恢復的技術。打個比方，就是將一段高清視頻壓縮成較小的文件，卻能在播放時基本保持畫質。還以餐廳檔口為例，過去檔口在和顧客交互中要記錄每一道菜，而現在把菜單中的菜品抽象匯總分類，計為十碗「例湯」、十碗「白飯」，要記錄的信息一下子就少了很多。

DeepSeek團隊在訓練V3版模型時，還使用了對數據資源節省技術，即「FP8混合精度訓練」。FP8是一種比常規大模型訓練使用的FP16和FP32更低精度的數據格式，每個數字佔用的比特數更少，這使得模型在訓練和使用過程中，所需的存儲空間和計算量大大減少，效率也就更高。之前大家也不是沒想過這種方式，但只有DeepSeek團隊真正實現了這一點。

專家 高飛

科技觀察家，至頂科技CEO兼總編輯，PEC China（提示工程峰會聯合發起人），長期研究人工智能等技術生態的發展規律，並參與了中國雲計算產業發展白皮書、中國科協學術場景AI技術應用評測報告等多個產業學術研究的立項編寫工作。



謬誤2

DeepSeek背後是接近50000台英偉達最先進A100芯片在支撐，不可能只用2000塊GPU就訓練出這麼高性能的模型。

高飛：否。DeepSeek的技術論文中清楚地寫明了訓練模型所需要的GPU數量，DeepSeek團隊沒有必要在這個數字上作假。目前所有聲稱「DeepSeek靠五萬卡訓練模型」的說法，都是坊間猜測，無任何權威出處。

謬誤3

DeepSeek模型只受國人追捧，美國有OpenAI、Meta等前沿模型，沒有必要用DeepSeek的產品。

高飛：否。近期全球眾多大廠紛紛宣布支持DeepSeek的模型，包括亞馬遜、英特爾、微軟、Perplexity、NVIDIA等。DeepSeek的APP也登頂多個國家的App Store下載排行榜。可以說，DeepSeek在全球範圍內都是一個現象級的產品。

Download



▼人工智能發展一日千里。圖為2016年3月9日至15日，人工智能程序「阿爾法圍棋」在韓國首爾進行的五番棋比賽中，以4比1的總比分擊敗韓國九段棋手李世石。新華社



強化學習 博採衆長

大公報：我們現在知道，DeepSeek的高性能根本上來自於新算法，即強化學習方法，這似乎和當年Alpha Zero與人類棋手博弈相似，是嗎？

高飛：是的。DeepSeek的強化學習，不是新理論。大家最耳熟能詳的強化學習模型，應該是谷歌DeepMind的圍棋模型Alpha Zero。它之所以叫Zero，是指模型是從零數據開始，通過與自身進行數百萬次對弈，積累數據，提升性能。

DeepSeek R1也是這樣，通俗地說，它就是人工智能界的「Alpha Zero棋手」，用AI和AI

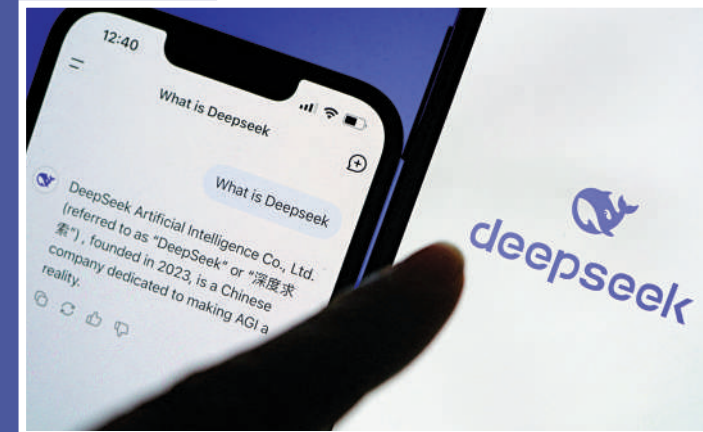
對弈的強化學習方式（而不是學習人類知識行為數據），提高性能。需要說明的是，DeepSeek R1並非單一地運用強化學習方法，而是新老方法並用，博採眾長。例如，DeepSeek團隊發現模型出現中英文雙語混用、答案不易被人理解的現象，所以也使用了傳統的監督式學習（SFT），即人類數據輔助優化，讓內容輸出更友好。

大公報：但谷歌DeepMind八年前就已推出圍棋模型Alpha Zero，強化學習法早已有之。在這方面，DeepSeek是否算抄襲了谷歌的技術呢？

高飛：這就是科學、技術、工程的區別了。強化學習技術早有公開論文，大家都可學習借鑒。但科學原理如何在技術和工程上實現，是另一回事。

此外，屬於封閉性問題（即有標準答案、可判斷勝負）的圍

棋問題，與處理開放性問題的語言大模型是不同的。這種強化學習技術，並不容易在大語言模型的訓練中使用，從去年開始湧現的千百個大模型，都沒能取得這方面突破，沒能走通這條技術路徑。



▲目前，DeepSeek下載量持續攀升，據估計，一周內DeepSeek網站用戶數量就超過了谷歌的Gemini聊天AI。