DeepSeek新模型開源 推理能力大升級

深度適配國產芯片 資金追捧相關股份

萬眾矚目 中國通用人工智能領先企業DeepSeek（深度求索）昨日發布大模型DeepSeek–V4預覽版並同步開源，該模型在Agent代理（智能體）執行力、世界知識儲備及邏輯推理性能上均取得顯著突破，標誌着國產開源模型正式跨入「百萬級超長上下文」普惠時代。

DeepSeek–V4分為「Pro」與「Flash」兩個版本，旨在滿足不同層次的應用需求。兩個版本在國內與開源領域處領先位置，而在數學、STEM、競賽型代碼的測評中，則取得比肩世界頂級閉源模型的優異成績。市場預期新模型將可以深度適配國產芯片，芯片股昨日大漲，華虹半導體（01347）急升15%。

大公報記者 劉鑛豪

深度求索今次推出兩款模型參數差距較大，DeepSeek–V4–Pro的參數為1.6T，而DeepSeek–V4–Flash的參數僅284B。深度求索以「性能比肩頂級閉源模型」形容DeepSeek–V4–Pro，主要體現在三個部分：Agent能力大幅提高、豐富世界知識、世界頂級推理性能。首先在Agent能力方面，相比前代模型，「Pro」的Agent能力顯著增強；在Agentic Coding評測中，達到開源模型領先水平。DeepSeek–V4成為公司內部員工使用的模型，據評測反饋，使用體驗優於Sonnet 4.5，交付質量接近Opus 4.6非思考模式，但思考模式存在一定差距。

「Pro」在世界知識測評中，大幅領先其他開源模型，僅稍遜於頂尖閉源模型Gemini–Pro–3.1。推理性能方面，「Pro」在數學、STEM、競賽型代碼的測評中，超越當前所有已公開評測的開源模型，取得比肩世界頂級閉源模型的優異成績。

支援百萬字超長文

至於「Flash」版本，深度求索形容為「高效經濟之選」。公司表示，雖然相比「Pro」，「Flash」在世界知識儲備方面稍遜，但推理能力接近。由於模型參數和激活更小，相較之下「Flash」能提供更加快捷、經濟的API服務。在Agent測評，「Flash」在簡單任務上與「Pro」旗鼓相當，只在高難度任務上有差距。

值得一提的是，DeepSeek–V4開創一種全新的注意力機制，在token（詞元）維度進行壓縮，結合DSA稀疏注意力，實現全球領先的長上下文能力，並且相比於傳統方法大幅降低了對計算和顯存的需求。兩款模型擁有百萬字超長上下文，同時支持非思考模式與思考模式，並成為深度求索所有官方服務的標配。該公司表示，用戶即日起登錄官網「chat.DeepSeek.com」或官方App，即可與最新DeepSeek–V4對話。

華虹與中芯抽高一成

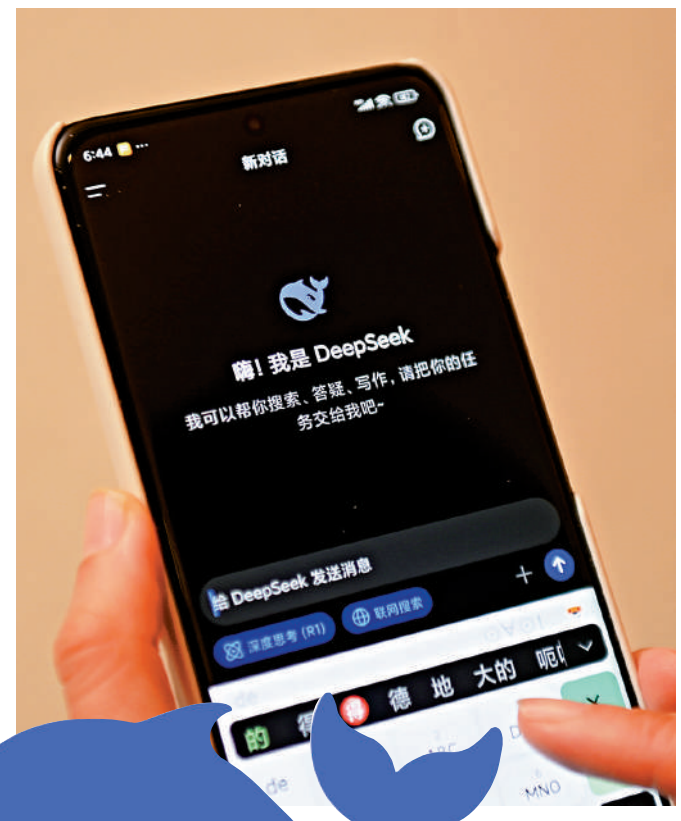
投資者預期DeepSeek–V4將適配國產芯片，芯片股昨日顯著造好，華虹半導體大升15.1%，收報108.1元；中芯國際（00981）升10%，報64.3元。豪威集團（00501）升7.3%，報87.75元。華贏東方證券研究部董事李慧芬認為，經過昨日拉升後，芯片股短線或回吐，投資者不妨待調整後再吸納。雖然芯片股中線可看高一線，以短線操作為主的投資者，需要緊貼股價走勢，因為股價波動較大。

加劇Agent賽道競爭

招商證券國際認為，因應DeepSeek–V4的推出，阿里巴巴（09988）、騰訊（00700）等龍頭雲廠商將直接受益，帶動MaaS平台收入持續提升。阿里巴巴、騰訊據報正洽談投資DeepSeek。

招商證券國際又稱，DeepSeek–V4–Flash輸出定價僅2元人民幣／百萬token，顯著利好AI應用層開發商成本效益，亦利好AI–native SaaS類公司加速AI能力內嵌。

該行相信，DeepSeek–V4對智譜（02513）、MiniMax（00100）等上市模型廠商不會構成顛覆性衝擊，但加劇Agent賽道競爭。



▲ DeepSeek–V4預覽版亮相，百萬字上下文成標配算力，顯存需求大降。

DeepSeek–V4亮點

- DeepSeek–V4–Pro、DeepSeek–V4–Flash最大上下文長度為1M（100萬），同時支持非思考模式與思考模式
- DeepSeek–V4–Pro的Agent能力顯著增強。在Agentic Coding評測中，V4–Pro已達到當前開源模型最佳水平
- 在世界知識測評中，DeepSeek–V4–Pro大幅領先其他開源模型，僅稍遜於頂尖閉源模型Gemini–Pro–3.1
- 在數學、STEM、競賽型代碼的測評中，DeepSeek–V4–Pro超越當前所有已公開評測的開源模型，比肩世界頂級閉源模型的優異成績
- DeepSeek–V4開創全新注意力機制，在token（詞元）維度進行壓縮，結合DSA稀疏注意力，大幅降低對計算和顯存的需求

芯片股昨日跑贏大市

股份	昨收（元）	升幅
華虹半導體（01347）	108.10	+15.1%
中芯國際（00981）	64.30	+10.0%
晶門半導體（02878）	0.40	+9.5%
天數智芯（09903）	457.00	+9.5%
納芯微（02676）	155.00	+9.0%
宏光半導體（06908）	0.49	+7.6%
豪威集團（00501）	87.75	+7.3%
ASMPT（00522）	165.20	+6.9%

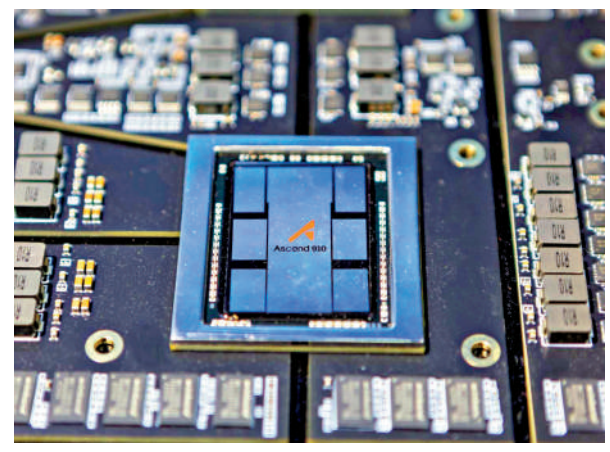
強強聯手 華為昇騰全面支持DeepSeek–V4

【大公報訊】伴隨DeepSeek–V4發布，華為表示，昇騰（Ascend）一直同步支持DeepSeek系列模型，今次通過雙方芯模技術緊密協同，實現昇騰超節點全系列產品支持DeepSeek–V4系列模型。昇騰是華為自研的基於「達芬奇架構」的AI處理器系列（NPU），專注於高性能邊緣與雲端計算。

華為指出，昇騰950超節點重新定義長文本推理的性能天花板，實現DeepSeek V4–Pro 20ms和DeepSeek V4–Flash 10ms低時延推理。DeepSeek承認，現時V4–Pro版本的API訪問服務吞吐十分有限，直到下半年昇騰950超節點批量上市後，價格將會大幅下降。

寒武紀昨日亦宣布，已基於vLLM推理框架完成對DeepSeek最新開源模型的適配，模型發布當日即實現穩定運行，適配代碼已開源到GitHub社區。

針對DeepSeek–V4的新結構，寒武紀通過自研高性能融合算子庫Torch–MLU–Ops，對Compressor、mHC等模塊進行專項加速；利用BangC高性能編程語言，編寫稀疏／壓縮Attention、GroupGemm等熱點算子的極致優化Kernel，充分釋放硬件底層性能。



◀華為昇騰通過雙方芯模技術緊密協同，支持DeepSeek–V4系列模型。

螞蟻旗艦模型 主打「快思考」降成本

【大公報訊】螞蟻百靈發布，面向即時任務執行的萬億級綜合旗艦模型Ling–2.6–1T。該模型依託MLA與Linear Attention的Hybrid架構創新，摒棄繁瑣的「慢思考」，以「快思考」機制實現高效推斷。Ling–2.6–1T僅憑極低Token開銷直達結果，極致壓縮輸出成本，旨在為即時任務執行提供高效模型解決方案。

談及Ling–2.6–1T的亮點，簡單而言，即「你要它做什麼，它就照做，而且做得特別快。」它用的是一套叫「MLA與Linear Attention的Hybrid架構」的技術，核心思路就一句話——摒棄「慢思考」，主打「快思考」，用最少的Token來完成。

分析指出，現在AI有個趨勢，大部分喜歡深度推理，當用家提出一個問題，它會一步步分析、拆解、驗證，最後給出答案，有些像人類做數學題的過程，雖然邏輯嚴密，但很費時。百靈Ling–2.6–1T走的不是這條路。它的思路是：許多日常任務根本不需要那麼複雜的推理，快速反應更重要。

騰訊新語言模型上線 增強與用家互動

【大公報訊】騰訊混元發布Hy3 preview語言模型，並直接開源，此模型總參數295B，活化參數21B。從體量來看，它算不上業界最大，但在複雜推理、上下文學習、程式碼、智能體等能力上均實現了大幅提升。同時，其賣點在於聚焦中等規模的模型最佳化，力求以更低成本實現高效AI應用。據悉，混元團隊為Hy3 preview明確立下了三條鐵則：能力體系化、評測真實性、追求性價比。總結下來，「一定要讓智能用得起、用得好。」

現時騰訊擁有微信、遊戲、廣告、雲端服務等極為複雜的業務場景，Hy3 preview自開發時便強調結合一線業務開展測評，其效果也有望更加貼近工作生活中真實的難點痛點。舉例說，混元團隊與元寶進行了深度的協同研發，由此產出的Hy3 preview能理解用戶在旅行過程中預算的突然變化、進而調整行程規劃，也能在提問者心情不好時不一味推出建議，而是重在寬慰，提供更智能且更具「活人感」的交互體驗。

Hy3 preview發布後，迅速在騰訊元寶、CodeBuddy、WorkBuddy首發上線，微信公眾號、QQ瀏覽器、QQ、騰訊新聞等十餘個主線產品也陸續接入。

花旗：定價具競爭力

花旗發表報告表示，騰訊的新模型Hy3預覽版，這是一款MoE（混合專家）架構語言模型，總參數量達2950億，上下文窗口256k，在複雜推理及代理能力方面有顯著提升。Hy3採取「務實」方針，專注於現實世界應用而非實驗室評分，並聲稱推理效率提升40%，定價具競爭力。

該行認為，儘管名為「預覽版」，但模型已立即整合至騰訊廣泛的產品線，包括元寶、ima、CodeBuddy、WorkBuddy、QQ、騰訊文檔、騰訊地圖、微信公眾賬號及和平精英等，顯示模型已準備好用於解決現實問題。模型專注於實用及具成本效益的部署，屬正面戰略方向。該行相信，隨着官方Hy3正式版或後續3.1/3.2版本推出，模型能力有望進一步提升。

▲騰訊混元發布Hy3 preview語言模型，並直接開源。

DeepSeek–V4兩款模型資料

模型	參數	激活	上下文長度	開源	API服務	網頁/APP訪問方式
DeepSeek–V4–Pro	1.6T	49B	1M	✓	✓	專家模式
DeepSeek–V4–Flash	284B	13B	1M	✓	✓	快速模式

逾十車企接入阿里千問 一句指令完成多項生活服務

【大公報訊】昨日是2026北京車展開幕首日，一則消息成為車展熱話。據內媒報道，十多家車企同一天宣布接入阿里千問。這代表什麼？日後相關車企的部分車款，駕駛員只需發出一句語音指令，系統便可規劃路線、查閱新聞、預訂酒店、購買演唱會門票，甚至點選外賣。一條指令，便能完成多項生活服務需求。分析指出，目前汽車市場競爭激烈，這些車企接入阿里千問大模型，有助強化產品差異化，爭取更多消費者青睞。

據了解，阿里雲提供了一整套「端+雲端」協同架構賦能汽車產業，打造體驗出色的智慧座艙。端側部署Qwen–Omni全模態大模型，既能感知周邊環境、保障用戶隱私安全，亦可在弱網環境下快速響應、穩定運作；雲端則透過千問連結數位世界，接入豐富多元的生活服務場景。

早前一汽紅旗品牌已率先將千問融入車載系統，並在紅旗HS6中正式搭載。分析認為，當下內地汽車市場競爭持續加劇，各大車企紛紛布局車載軟體與智慧服務，藉此建立差異化優勢吸引消費者，千問車載版亦順勢迎來大規模普及。

另有分析提到，十多家車企集體接入千問，覆蓋自主、合資、新勢力多個板塊，正式形成「阿里系」座艙陣營，勢必對華為帶來明顯競爭壓力。過去車企布局車載智能化，首要合作對象往往是華為。如今千問同步牽手十多家主流車企，意在向市場證明，華為能夠實現的智慧座艙功能，千問亦有力提供相關技術和服務，且車型覆蓋更廣，預計未來阿里與華為在車載AI領域的正面比拼，將愈趨頻繁。

車載AI從「聽指令」變「辦實事」

市場人士指出，千問的核心賣點不在單純的技術參數比拼，其真正競爭優勢，在於可調動整個阿里生態資源，讓車載AI由單純「聽懂指令」，升級為主動「辦成實事」。例如駕駛員發出複雜指令：「先去接孩子，然後找一間沿途評分較高的川菜館，預訂晚上七點的位置」，千問可自動拆解任務邏輯，分步完成導航、搜店、訂位等一系列操作。隨着千問加速落地普及，將倒逼華為持續強化系統流暢度與硬體外，亦需加大生態服務的布局，擴展服務廣度與深度，應對這一強勁對手帶來的挑戰。

◀阿里千問獲十多家車企接入，形成阿里系座艙陣營。