大公報 2026年6月23日 星期二

算力成本膨脹 企業AI代人手得不償失

專家：處理任務越繁瑣複雜 消耗詞元越多

經濟透視

人工智能（AI）技術持續發展，企業競相推動員工擴大使用AI以提升生產力，但它眞的能大規模取代人力並產生更高經濟效益嗎？近期較多人談論的話題，是AI成本暴漲，已遠超企業的預期。早前有報道指出，美國叫車與外送平台優步（Uber）在今年頭四個月內，因工程師追求以Token（詞元）為中心的績效，已耗盡2026年一整年的AI預算，凸顯出企業將AI整合至營運的實際開銷過大。

面對如此高昂的支出，企業或許需要思考，「透過裁員改用AI就能大幅節省成本」是否只是一個幻想？有機構研究指出，現階段盲目將AI投入所有工作流程中不但無法獲利，反而會因為高昂的算力成本而導致經濟效益不划算。

大公報記者 李耀華

◀研究指出，即便Token單價預計在2030年前下降近九成，企業AI總成本仍可能持續攀升。

優步技術總監Praveen Neppalli Naga早前證實，從年初至四月間的短短數月，便已耗盡了一整年的AI預算，原因是要向大約5000名工程師發給Anthropic的Claude編碼，燒錢速度遠超出原來財務模型所預計的。

四個月花光一年詞元預算

新聞網站《Axios》日前指出，Uber在今年頭四個月內，就將2026全年度的代理型AI（如Anthropic產品）使用預算全部燒光，凸顯出將AI整合至營運的實際開銷規模之大。事件的導火線，源於科技界近期出現的新現象Tokenmaxxing（Token極大化）。這個現象指的是企業或開發者不顧成本與效率，盲目追求讓AI處理或生成最大數量的Token，企圖以極端手段來提升自動化表現。

Token降價 難抵銷用量暴增

Token是AI處理資料的基本單位，以Token為基礎的定價系統是當今AI與大型語言模型（LLM）的主流計費模式。Uber內部向旗下5000名工程師全面開放ClaudeCode與Cursor等自主型AI工具，並設立了內部的「Token消耗排行榜」，用以衡量開發進度。在工程師眼中，Token消耗量愈高，代表愈積極使用AI，排行愈前。結果，工程師展開了瘋狂的刷數據比賽，放任AI自動化工具在背景無休止地運作。每名工程師每月的API Token開支，急升到500至2000美元，導致全年AI預算在四個月內徹底清袋。

Uber在去年的整體研發開支達34億美元，按年升9%。然而，今次令該公司提早耗盡AI預算開支並不是因為財政規模的問題，而是Uber的財務部門還未有學懂如何管理好新的定價模式。事實上，以Token為基礎的定價系統與以軟件為基礎的不同，財務總監懂得如何為後者建立預算模型，但前者卻未能夠這樣，因此工程師的開支與財務部本身的預算便有了極大落差。

高盛早前預測，隨着企業與消費者廣泛採用AI代理，到2030年，AI代理可能使Token消耗量暴增24倍，達到每月高達120千萬億Tokens的驚人規模。研究機構Gartner也警告，即便Token單價預計在2030年前下降近九成，企業AI總成本仍可能持續攀升，因為AI代理執行每項任務所需的Token量，遠超傳統模型，使用量的成長速度可能輕易抵銷降價帶來的紅利。

這些信號顯示，以AI取代或輔助人力的經濟效益，遠比早期預測來得複雜。當Token消耗增速持續超越單價降幅，那個由AI代理驅動的企業未來，恐將帶來遠超行政總裁預期的龐大賬單。

人工智能開支龐大，不僅反映在Uber的開支上，還可以從英偉達的高層口中得知。英偉達應用深度學習副總裁向《Axios》表示，就其團隊而言，人工智能「運算成本遠遠超過了員工成本」。這項說法進一步印證了AI部署對企業財務構成巨大挑戰。

MIT：視覺能力崗位77%靠人

整體而言，研究顯示，即使AI系統能夠正確部署，其所需成本也幾乎總是高於人類勞工。這使得AI大規模取代人力的策略在經濟可行性上，仍有待更深入的評估。然而，所謂AI取代人力的論點不能一概而論。有時AI系統不僅未能省下人力成本，反而可能讓企業開支超出當前人力所需的成本。麻省理工學院（MIT）2024年一項研究的觀點顯示，僅在視覺能力為核心的23%職位中，AI自動化才具有經濟效益，其餘77%仍依賴人力。

◀Uber去年整體研發開支達34億美元，按年升9%。

▲迪士尼投資78億元，允許Sora用戶使用迪士尼角色生成影片，但OpenAI於3月宣布退出影片生成業務，反映視頻生成面對巨大競爭。

企業AI開支情況

❶ 優步今年頭四個月已耗盡一整年的AI預算

❷ 高盛預測，到2030年，AI代理可能使Token消耗量暴增24倍，每月高達120千萬億Tokens規模

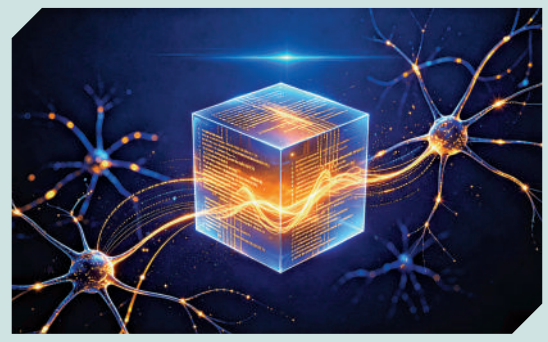
❸ 英偉達稱，人工智能運算成本遠遠超過員工成本

❹ Gartner警告，即便Token單價預計在2030年前下降九成，企業AI總成本仍可能持續攀升，因為AI代理執行每項任務所需的Token量，遠超傳統模型

❺ 據估算，以Sora模型生成高質影片，背後需要龐大的運算資源，折合每秒鐘高達數百港元，OpenAI早前宣布退出影片生成業務

❻ 早前在硅谷廣泛流傳的消息指，某家企業在Claude上的一個月支出，高達5億美元

❼ 有團隊做Multi-Agent（多AI協同）實驗，僅一晚上，Token費用直接突破百萬人民幣



▲全球AI大模型的價格正受惠於行業價格戰而大幅降低。 AI製圖

DeepSeek旗艦模型減價 爭企業級AI服務市場

競爭優勢

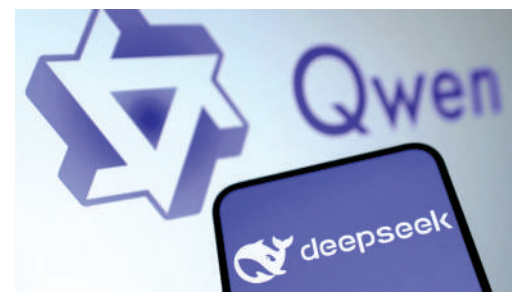
這邊廂美國企業為着AI而花費巨資，但那邊廂中國的AI成本卻不升反跌。中國AI公司DeepSeek近日宣布，旗下旗艦模型V4Pro永久降價75%，大幅降低企業使用成本。此舉不僅對OpenAI和Anthropic等美國主要AI業者構成直接競爭，更預示着企業級AI服務市場將迎來結構性變革，促使全球企業加速評估更具成本效益的AI解決方案。

DeepSeek V4 Flash模型在輸入成本上，比Anthropic的Claude Sonnet或者OpenAI的GPT5.5–Med便宜7倍，輸出成本更便宜17倍。此外，輕量級的DeepSeek V4 Flash模型，相較於Claude Haiku等入門級替代方案，價格是原來的十分一甚至二十五分一。

DeepSeek能夠實現如此大幅度的降價，主要原因是其在軟硬體整合上的創新。早在2024年的V2架構中，DeepSeek就已透過序列維度壓縮、原生記憶體卸載等四項突破性技術，大幅降低模型運行所需的記憶體與運算資源。例如，DeepSeek V4 Flash僅需5.48GB的高頻寬記憶體（HBM）即可處理一百萬個Token的上下文，相較之下，其他西方小型模型可能需要高達89GB。這項技術不僅提升效率，也被視為規避美國對英偉達頂級GPU出口限制的地緣政治策略。

用家看重服務商收費

這次降價策略已開始影響市場動態。DeepSeek V4 Flash模型在OpenRouter排行榜上奪下首位，其Token使用量激增48%，而DeepSeek V4 Flash也名列第六。包括Uber、Airbnb和Pinterest在內的企業，都曾因高昂的Token使用成本而尋求替代方案。Uber一名主管表示，缺乏更好的產品展示，高昂的Token費用越來越難以證明其合理性；Airbnb行政總裁Brian Chesky則傾向使用阿里巴巴的Qwen等更快速、便宜的模型。Pinterest技術總監Matt Madrigal透露，該公司透過在專有資料上訓練Qwen模型，將成本降低了九成。



▲DeepSeek與阿里巴巴的Qwen等國產大模型具價格競爭優勢。

輸入指令與生成內容 均計作詞元

技術拆解

Token是AI模型的最小單位，當用家在使用ChatGPT、Claude或Gemini時，每一次輸入與回覆，背後都在消耗一種看不見的單位「Token」。

簡單來說，Token就是AI處理文字時的最小單位。在大型語言模型（LLM）中，所有輸入的句子都不會直接被「閱讀」，而是會先被拆解成一個個Token，接着再轉換成對應的數字編碼，模型才能進行後續的運算與預測。

換句話說，人類看到的是完整語句，但AI真正處理的是「Token＋數字」。這也是為什麼AI的核心運作並不是理解語意，而是透過大量資料訓練後，去預測下一個最可能出現的Token，進而產生看似合理的回答。

因此，當內容越長、對話越多，Token就越多、成本也越高。在多數AI服務中，Token不只是資料處理單位，同時也是計算費用的基礎。無論是輸入的內容，還是模型產生的回覆，都會被計算成Token，並轉換為使用成本。

這就是說，當輸入越長的指令，或是讓AI產生越多內容時，所消耗的Token就會越多，相對的費用也會提高。對於使用API（應用程式界面）或進階方案的使用者來說，這種影響會更加明顯，因為每一次互動背後都對應到實際的資源消耗。而所謂API，簡言之，是指可以幫助開發者節省精力，並很快地達到目的的界面。

模型內部推理仍產生成本

另外，在某些進階模型中，還可能存在額外的Token類型，例如用於加速回應的快取Token，或是模型內部推理過程所使用的Token。雖然一般使用者不一定會直接感受到這些差異，但它們會影響整體成本與效能。

以ClaudeCode為例，是Anthropic推出的一款代理型（Agentic）AI工具，能控制用家的電腦終端機，擁有查看檔案、編輯程式碼、執行指令的權限，故所使用的Token也會更多，成本更高。

▲由Anthropic所開發的AI助手Claude，可以直接在電腦上執行多個步驟任務。

AI影響勞工市場正擴散

財經觀察 李耀華

美國科技巨頭在近年加緊裁員，目的是要集中資源發展人工智能。不過，隨着AI日益普及，裁員的公司已不再局限於科企，而廣及金融機構。本來，企業為了增加競爭力而加強在AI的應用無可厚非，但是，愈來愈多研究報告指出，企業花費巨資發展AI可能最終並無明顯好處，反而因削減成本裁員會影響公司商譽。

高盛經過兩年時間對生成式AI進行研究後，發現公司即便花了巨額資本，但實際效用仍然不大，換言之，投資與回報仍然存在落差。雖然向AI傾注了數百億美元，但多達95%的公司仍未見到其投資有明顯的回報。

報告指出，雖然AI在客戶間的普及速度比當年的互聯網和個人電腦要快，但實際上卻未能令企業的利潤有顯著增長。原因是多數用家都依賴免費的AI工具，令AI供應商的營業額增長受到限制。

另一方面，企業為了趕AI浪潮，都作出了龐大的投資，麥肯錫與研究機構Gartner的預測顯示，AI基礎建設將持續推升全球IT支出。Gartner預估，2026年全球IT支出將高達6.31萬億美元，年增13.5%。同時，根據摩根士丹利統計，科技巨企持續大舉投資AI，光是今年在這方面的資本支出就高達7400億美元，較去年增長69%。

美首季9萬人遭解僱

面對不斷擴大的開支，企業唯有開源節流，在預計AI將會取代某些人手的情況下，斷然展開大裁員行動。據職場狀況追蹤網站Layoffs. fyi的統計，美國今年首季有超過9萬人遭到解僱，今年全年裁員規模勢將超過去年全年549家公司的逾15萬裁員數字。

裁員數字不再局限於科企，金融界亦開始受到影響，例如摩根大通行政總裁戴蒙表示，隨着AI普及，集團將減少傳統銀行家的招聘，轉而招募更多AI專家。而渣打銀行更計劃在2030年前，透過擴大採用AI與自動化技術，在全球裁減約7800個後台營運與企業職能崗位，佔其總員工數近10%。

然而，如果這些企業最終發現AI實際上的效益未如理想，外界便可能會質疑裁員行動是否合理，到最後更會損害到有關公司的商譽。

◀愈來愈多研究報告指出，企業花費巨資發展AI可能最終並無明顯好處，反因削減成本裁員而影響公司商譽。 AI製圖