

3C科技 今年3月，國家數據局在中國發展高層論壇上正式宣布，將Token的中文標準譯名定為「詞元」。Token指的是人工智能模型用來理解和處理語言的基本文本單元。這個曾經只屬於計算機科學教科書的術語，終於有了一個官方的中文名字。名字的落定，恰恰標誌着一個更重要事實的確認：Token正在成為AI時代衡量成本、效率與價值的核心尺度。以下通過當前科技企業對Token的運用，討論Token將如何改變你我的生活。

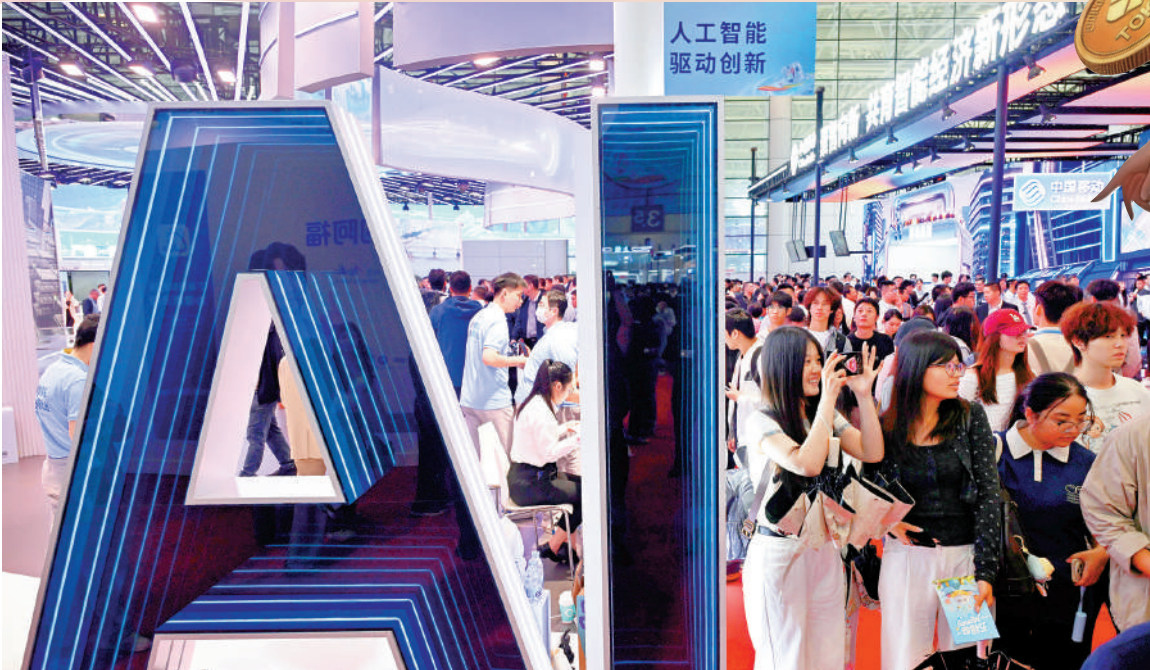
姚剛

AI時代衡量成本、效率與價值的標尺



掌握Token 高效使用砵碼

速場。AI應用向智能體升級，帶動詞元（Token）消耗快速攀升。圖為4月29日，第九屆數字中國建設峰會現場。



人工智能
驱动创新



大公報AI製圖

從「隨便用」到「算着用」

英偉達CEO黃仁勳在今年年初拋出一個極具衝擊力的觀點：「如果一位年薪50萬美元的工程師，消耗的Token價值還不到25萬美元，我會非常擔心。」這番話在硅谷掀起了連鎖反應。黃仁勳透露，英偉達正努力為工程團隊投入約20億美元用於AI Token算力額度，並考慮將Token預算納入薪酬體系。他在今年的英偉達GTC大會上提出了一個核心判斷：Token是AI時代的新大宗商品。未來的數據中心將演變為生產Token的工廠，其商業邏輯由計算效率與產出價值直接驅動。黃仁勳表示，隨著模型參數規模擴大、上下文窗口延長，AI的智能程度會持續提升，但詞元的生成速率會隨之下降。在這套體系中，誰的工廠效率更高，誰就能在下一輪競爭中勝出。

隨着AI深入職場，Token正在變成衡量人力價值的第四維度——薪資、獎金、股權之外，再加一項「算力配額」。據傳，有科技巨頭開始推行「Token消耗量考核制度」。然而，這種量化管理很快暴露了它的另一面。Uber在去年12月為5000名工程師配發了AI編程工具，結果到今年4月，全年的AI預算就用光了。無獨有偶，Meta內部測算顯示，若維持當前員工調用增速，2026年僅內部AI使用一項支出便將高達數十億美元。今年6月之前，騰訊部分部門給員工的AI使用額度一度達到每人每月約2000美元。調整後，不少員工每月只剩1000至5000元人民幣的Token預算，有員工感慨「兩天就用完了」。從「隨便用」到「算着用」，這個轉變只用了半年時間。Token作為生產力標尺的使命，才剛剛開始就被自己的刻度反噬了。

由此可見，未來的職場精英不僅僅是會寫代碼或設計界面的人，更是懂得如何高效利用AI算力、如何在有限的Token預算下實現最大產出的人。Token消耗效率，將成為衡量專業人才核心價值的關鍵指標之一。

業內差異化的商業模式

Token作為AI時代的核心流通商品，催生了差異化的商業模式。當前行業內各大玩家基於自身技術和資源優勢，形成了免費引流、生態綁定、付費變現、算力中轉等四種主流商業模式，不同路徑背後，既是企業的戰略布局，也是行業競爭格局的真實縮影。

用免費搶用戶

今年7月，模型廠商Agnes AI推出開源編碼模型Agnes-2.5-Flash，以完全免費的模式顛覆了

AI編程行業的定價規則，成為行業現象級事件。其模型能力雖相較於頂級付費模型仍存在一定差距，但在代碼編寫、漏洞修復、多文件協同開發等真實落地場景中，性能差距幾乎可以忽略，完全滿足工業級開發需求。

它最具顛覆性的是推出無限制免費服務模式：API調用永久免費，無次數、速度、時長限制。數據顯示，模型上線7天內，平台累計處理5.41億億詞元算力請求，成功經受住了生產級超高併發流量的考驗。免費的Agnes-2.5-Flash出現，讓全球開發者能夠零Token成本享用工業級AI算力。

補貼換生態

幾乎同一時間，騰訊雲宣布微信AI小程序成長計劃：面向全行業開發者免費疊加10億大模型Token額度，AI生圖免費額度從1萬張暴漲至10萬張，底層模型全面升級至混元Hy3。10億Token按市場價折算價值不菲，但騰訊的算盤其實很清楚——用算力補貼換取生態黏性，讓開發者在微信生態裏扎根。依託微信13億月活的頂級流量池，開發者一旦適配騰訊雲開發體系、熟悉混元模型調用邏輯，將形成極高的遷移成本，最終實現開發者、模型、平台的深度綁定。



▲2026中國國際大數據產業博覽會以「詞元——數據要素價值釋放新路徑」為主題。圖為小朋友在數博會上與機器人互動。

新華社

被迫的收費

字節跳動則是另一個故事：去年資本開支約1500億元，大部分集中在AI；今年已上調至2000億元。當利潤被持續侵蝕，商業化成了不得不走的路。6月的第三周，豆包正式推出專業版付費訂閱，消息一出，爭議迅速席捲社交媒體——兩個月前9.1萬人參與的投票中，8.8萬人反對收費。付費用戶的反饋也並不樂觀：額度用得太快像開盲盒，

▼Agnes AI依靠免費的模式吸引用戶。



完全不知道做完一個任務會消耗多少。

儘管豆包官方多次承諾基礎功能永久免費，但互聯網行業的發展規律讓用戶難以打消顧慮。過往大量互聯網產品的迭代歷程證明，付費體系落地後，免費版本的資源配額、功能權限、響應速度會逐步壓縮。

AI時代「賣鏟人」

有人則選擇了一條更中立的路徑——做AI時代的「賣鏟人」，只做算力的統一調度和分發。這條賽道上最受矚目的中國玩家是「硅基流動」。這家成立僅3年的公司，不造大模型，不做應用，只做一件事：把英偉達、華為昇騰等不同廠商的芯片通過自研引擎統一調度，讓DeepSeek、Qwen等模型跑起來，再按Token量賣給開發者和企業。去年其營收5533萬元，同比增長653%。但招股書同時顯示，毛利率從2024年的39.4%驟降至-24%，淨虧損3.45億元。每賺1塊錢倒貼2毛4——Token工廠的生意，遠沒有看上去那麼風光。

Token越多，價值越少？

去年硅谷流行一個詞叫「Tokenmaxxing」——鼓勵員工盡可能多地使用AI。到了今年，畫風突變：網絡安全企業Palo Alto Networks的CEO最近接受採訪時表示，Token價格還需要再下降90%，企業才能真正大規模部署；軍工企業Palantir的CEO則更直接批評：如果AI真能幫企業多賺10億美元，模型公司完全可以拿走其中30%，而不是按Token收費；研發管理平台Jellyfish的調查顯示，使用Token最多的工程師，生產力大約是低使用者的2倍，消耗的Token卻是後者的10倍。

為什麼AI這麼燒錢？因為Agent和傳統軟件的運行方式完全不同。傳統軟件執行路徑固定，成本可預測；而Agent會自主搜索、調用工具、反思、重試，整個執行過程充滿不確定性。有研究發現，一個Agent完成同類任務消耗的Token可能是普通代碼問答的1000倍。更諷刺的是，模型準確率通常在中等等級Token消耗區間達到峰值，之後便進入明顯的收益遞減階段。

Token天然適合計費，卻不天然代表價值。這可能是整個AI產業最根本的矛盾。

沒有贏家的遊戲

費」的博弈撕裂着商業模式——我們必須追問：Token經濟的終極形態，究竟是一場普惠的技術革命，還是一場風靡全球的數字狂歡？

Token本身沒有善惡，它只是鏡子，映照出我們選擇如何使用技術、如何分配價值、如何定義進步。對於個人而言，掌握Token高效使用邏輯，是適配AI職場、提升核心競爭力的關鍵；對於企業而言，建立Token精細化管控體系、平衡算力投入與產出，是降本增效、構

築競爭壁壘的核心抓手；對於整個行業而言，Token定價體系的完善、商業模式的成熟、算力資源的普惠均衡，將決定AI產業的發展上限。

每一次微小技術單元的定義重構，都是一次產業變革的前奏。未來，誰能掌控Token的效率密碼，誰就能在AI產業的新一輪浪潮中，搶佔先機、打開格局，解鎖人工智能真正的產業價值。

在中國，移動互聯網時代培養了消費者對「免費」的潛意識期待，這讓Token付費訂閱模式舉步維艱。即使在付費訂閱盛行的歐美，OpenAI也遠未實現盈利。去年，OpenAI收入130.7億美元，總支出340億美元，淨虧損385億美元。一家估值超過8000億美元的公司，至今仍在靠燒錢維繫增長。

Token正在走向零毛利。Token未來將成為「水、電、網」一樣的基礎設施，靠賣Token本身可能永遠賺不到大錢，真正的商業價值在於Token之上構建的服務和應用。這讓人想起一個古老的寓言：淘金熱裏賺最多錢的，從來不是淘金者，而是賣鏟子的人。但今天的AI淘金熱裏，連賣鏟子的都在虧錢。

Token從一個技術術語變成經濟單元，再變成成本黑洞，只用了不到3年。它既是AI時代的通行貨幣，也是這個時代最昂貴的實驗品。我們正在用真金白銀回答一個還沒有人知道答案的問題：智能，到底值多少錢？也許答案不在Token的單價裏，而在Token創造的價值裏。



▲豆包正式推出專業版付費訂閱引發網絡熱議。

寫在最後

Token經濟的崛起，本質上是人類第一次將「智能」本身變成了可計量、可定價、可交易的商品。這是一個劃時代的轉變——知識從人力資本轉化為數字資產，實現了模塊化生產與自動化交付。然而，一方面Token讓AI能力變得觸手可及，小團隊可以憑藉極低成本與大廠同台競技；另一方面，當Token消耗成為新的成本中心，當企業發現投入與產出嚴重錯配，當「免費」與「收