

行業概覽

本節及本文件其他章節所載資料及統計數據，摘錄自不同的官方政府刊物、公開市場研究的可得資料及其他獨立供應商來源，以及弗若斯特沙利文就[編纂]編製的獨立行業報告（「弗若斯特沙利文報告」）。來自官方政府來源的資料未經我們、聯席保薦人、[編纂]、彼等各自的任何董事及顧問，或參與[編纂]的任何其他人士或各方獨立核實，且對於該等資料的準確性亦無發表任何聲明。

資料來源

我們委聘獨立市場研究顧問弗若斯特沙利文就詞元供應平台市場進行分析，並編製弗若斯特沙利文報告。弗若斯特沙利文為一家獨立全球諮詢公司，於1961年在紐約成立，其服務涵蓋（其中包括）行業諮詢、市場戰略諮詢及企業培訓。我們已就所獲得的市場研究服務向弗若斯特沙利文支付人民幣550,000元的費用，董事認為該費用與市場價格相符。弗若斯特沙利文在編撰及編製弗若斯特沙利文報告時進行了以下調研：(1)一手研究，包括對行業參與者、競爭對手、下游客戶及業內權威第三方協會進行訪談；及(2)二手研究，包括審閱企業年度報告、相關官方機構數據庫，以及弗若斯特沙利文過去數十年積累的專屬數據庫。弗若斯特沙利文報告中的市場預測基於預測期內的以下主要假設：(i)預測期內所涉中國市場的社會、經濟及政治環境將維持穩定；(ii)預測期內中國詞元供應市場的相關政策將保持連貫一致；及(iii)詞元供應市場將受弗若斯特沙利文報告中所述因素的推動。本節所載全部數據及預測均來源於弗若斯特沙利文報告。該委託報告由弗若斯特沙利文獨立編製，未受本公司或其他利益相關方的干預。董事確認，經作出合理查詢後，就彼等深知，自弗若斯特沙利文報告日期以來，市場信息並無任何可能對本節所載資料構成限定、抵觸或影響的重大不利變動。

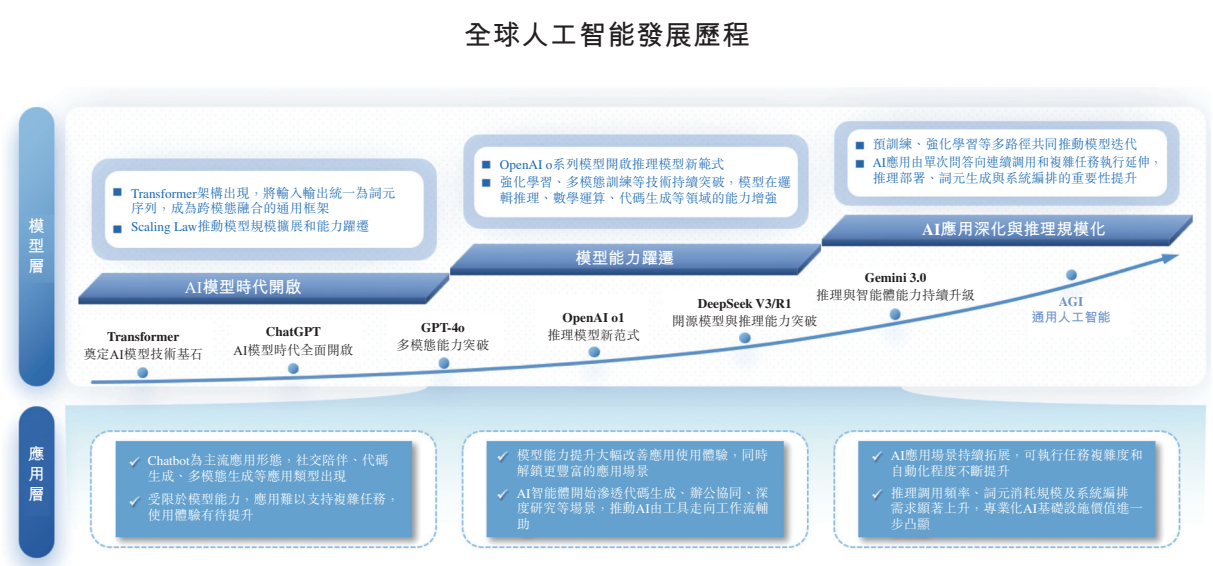
第四次工業革命的背景下，AI產業的基礎設施層價值分析

AI正在成為新一輪工業革命和生產力革命的核心驅動力，繼蒸汽、電力、信息技術之後，推動全球經濟社會進入以智能化為重要特徵的發展新階段。從市場規模看，按收入計，全球AI市場已由2021年的約人民幣2.2萬億元增長至2025年的約人民幣5.9萬億元，年複合增長率為27.6%；預計將以36.0%的年複合增長率增至2030年的約人民幣26.3萬億元。

相較於以往歷次工業革命，AI技術迭代更快、應用擴散更廣、對各產業的滲透與帶動更強，不只是提升效率，而是在系統性地重塑生產力體系、商業形態與公司組織架構。歷次工業革命都依托一套基礎設施：鐵路之於蒸汽時代、電網之於電氣時代、通信網絡之於信息時代；本輪亦然——支撐模型訓練、推理運行、算力資源編排與服務交付的AI基礎設施，正是其價值得以兌現、規模得以擴張的底層核心支柱。

行業概覽

AI產業發展歷程及其特點



資料來源：弗若斯特沙利文

- **AI產業仍處中早期，算力和AI模型率先受益；長期隨著應用大規模持續落地，系統軟件價值將持續放大**

AI產業主要由應用層和基礎設施層構成。應用層面向海量用戶和各類細分行業場景，提供各類AI智能體、軟／硬件產品及智能化解決方案；基礎設施層則包括算力、AI模型和系統軟件等主要環節。

當前AI產業處於中早期，算力和AI模型作為AI能力生成、部署和調用的關鍵底座，具備更強確定性，率先獲得商業和資本共識。長期來看，隨着AI應用大規模商業化的加速，上層應用價值逐步凸顯，將帶來更大規模的商業機會。在此進程中，系統軟件的價值將隨著應用數量、調用頻次、模型複雜度和企業級部署需求等方面的提升而持續放大——一方面向下持續提升異構算力和各類AI模型的利用效率，另一方面向上支撐應用側對高併發、低延遲、穩定交付和綜合成本優化的需求，成為承上啟下的關鍵層。

- **大規模部署AI應用和價值兌現推動算力需求從訓練加速向推理遷移**

訓練與推理是AI能力形成與價值兌現的兩個主要環節。其中，訓練決定模型基礎能力，通常是模型能力形成階段的集中式投入；推理則發生於AI模型的實際使用環節，是決定模型能力能否面向真實用戶、真實場景和真實業務持續交付的價值兌現環節，具有高頻、持續、分散、實時的特徵。

行業概覽

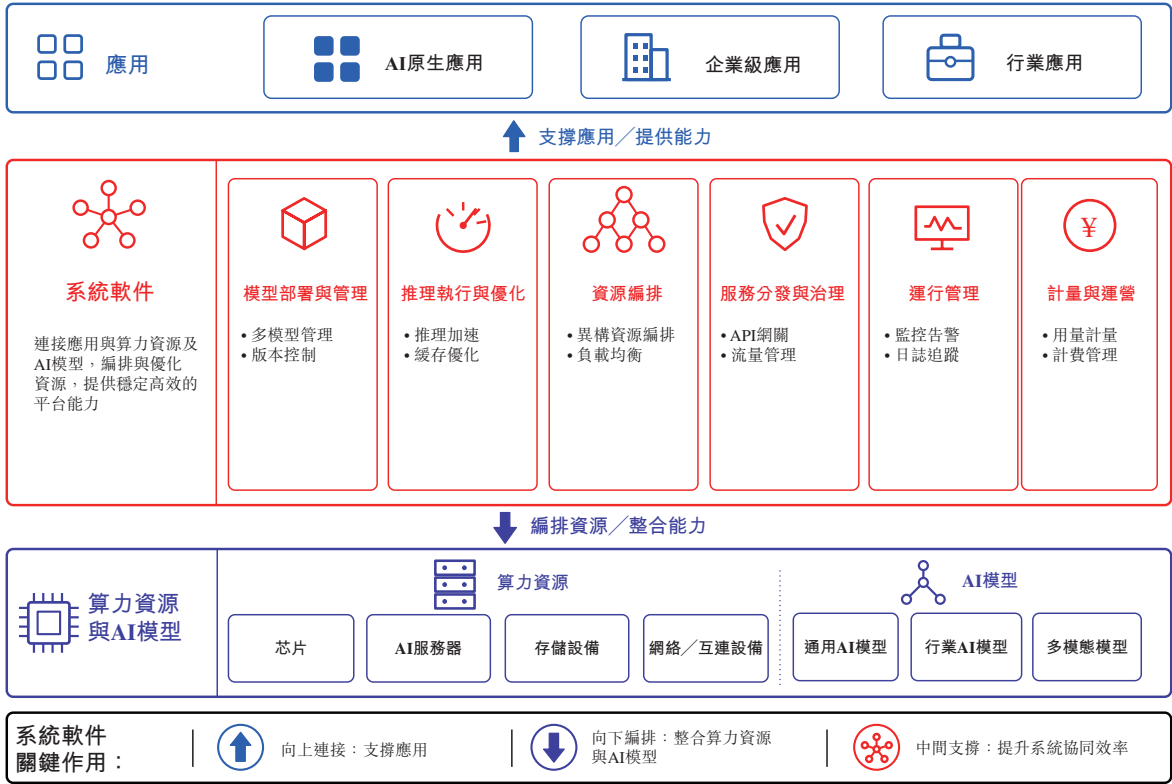
隨著實際場景需求爆發，AI聊天機器人、代碼生成與軟件工程助手、企業知識庫、文生圖／視頻創作工具、自動駕駛解決方案及各類智能化應用持續成熟，AI應用正從單次問答、文本生成等基礎場景，向多模態交互、複雜任務執行和連續智能服務延伸，並相應產生更高頻的用戶交互、更長的上下文處理、更強的實時響應需求和日趨攀升的任務複雜度，大幅推高AI模型的推理需求和詞元調用規模，推動AI產業的算力資源重心由訓練側加速向推理側遷移，單一模型在其推理階段內被消耗的詞元總量，將遠超其訓練階段。

AI基礎設施層的要素構成及系統軟件的關鍵作用分析

- AI基礎設施層主要包括算力、AI模型及系統軟件

AI應用的發展依賴基礎設施能力的持續提升。AI基礎設施層主要包括算力、AI模型及系統軟件。算力包括AI芯片、AI服務器、存儲設備及網絡/互連設備，為模型訓練及推理提供底層資源基礎及運行環境。AI模型包括通用AI模型、特定行業AI模型及多模態AI模型，決定AI能力的邊界及應用範圍。系統軟件作為組織及利用算力的中間層，承擔模型部署、推理執行、異構算力資源編排、服務分發等功能，從而使算力及AI模型潛力最大化。

系統軟件的關鍵作用



資料來源：弗若斯特沙利文

行業概覽

- **詞元供應平台已成為AI基礎設施系統軟件的重要產品及商業形態**

透過適配及封裝算力與AI模型，系統軟件為下游應用提供可調用、可管理及可持續運行的AI能力。在AI應用規模化落地及推理需求快速增長的背景下，行業內湧現了詞元供應平台這一構建於AI基礎設施系統軟件之上的產品及商業形態。詞元供應平台整合了推理引擎、異構算力編排、模型部署、服務分發、營運管理、計量及詞元交付等關鍵系統軟件能力。透過圍繞詞元生產、編排、優化及交付來組織及協調該等能力，詞元供應平台將分散的算力資源及模型轉化為穩定、高效及可擴展的詞元服務。

- **詞元供應平台成為跨芯片適配、模型部署和推理效率提升的關鍵價值環節**

目前，AI推理的特點是在芯片架構、算力、模型類型、部署方式或客戶需求等多個方面，有多種方案存在。同時，芯片及AI模型持續推陳出新，進一步推升在推理層面的快速適配、遷移及性能優化的要求。算力側，不同芯片在硬件架構、軟件生態與適配成熟度上各有特點；模型側，不同模型在參數規模、上下文長度、推理效率與應用上各有側重，且AI模型、通用與垂直模型迭代快速；客戶側，企業在無服務器詞元服務、專屬實例、本地部署等方面形成差異化訴求。

這種多元格局推動了產業鏈的專業化分工，也顯著抬高了"跨芯片 × 多模型"組合在不同類型客戶和應用場景的規模化落地的複雜度。因此，作為詞元供應平台的系統軟件層的核心價值也逐漸凸顯：一方面，適配和封裝多種算力芯片和模型，完成算子、精度、芯片內存與推理引擎的深度調優，從而為不同的應用場景提供穩定的推理服務。另一方面，通過推理引擎優化、算力資源編排、服務及營運管理等手段，在既定硬件上持續提升單位算力的詞元生產效率，並保障服務的穩定性與可規模化交付。

因此，隨著AI應用規模化落地，面向推理場景的AI基礎設施系統軟件需求將持續增強。具備跨芯片、多模型、多場景適配能力的詞元供應平台，將成為支撐AI產業應用規模化落地，乃至價值釋放的關鍵基礎設施。

AI應用爆發背景下的詞元革命及其深遠影響

- **詞元正由模型調用的計量單位，演進為面向終端應用的標準化產品形態**

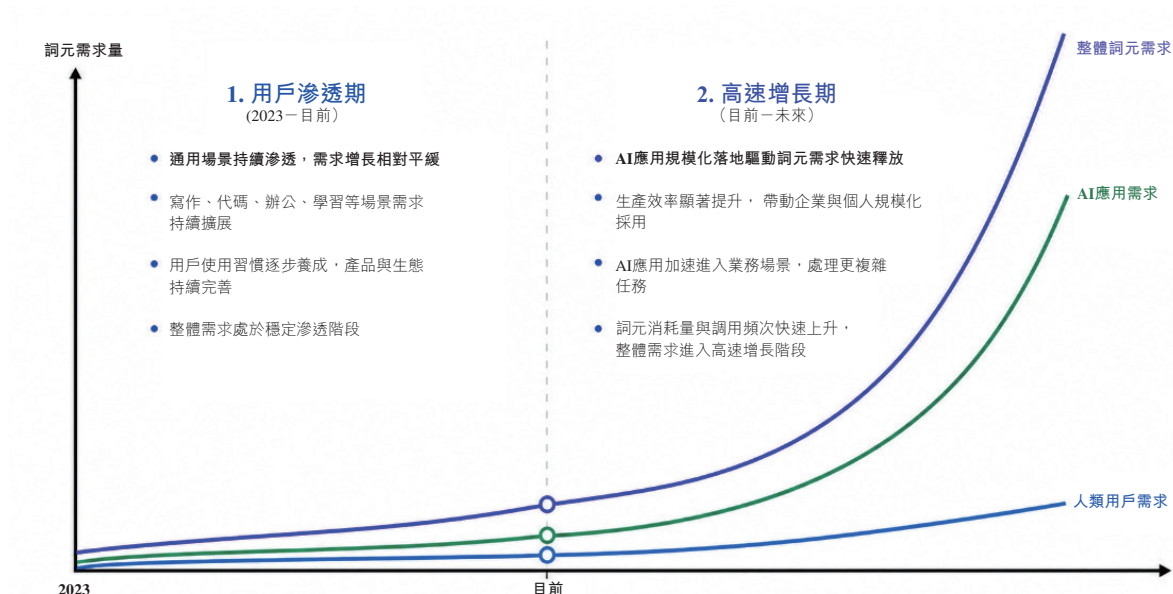
AI應用正由單次問答向多模態互動、複雜任務執行及連續智能服務演進，對應更高頻、更連續的詞元服務需求。2025年，中國詞元調用量快速擴張達到約2,426.3萬億詞元。詞元不再只是AI模型輸入輸出過程中的計量單位，而是逐步成為面向終端應用的標準化產品形態：詞元供應平台通過算力、AI模型及系統軟件能力生產和交付詞元，AI應用則通過持續調用和消耗詞元實現智能服務。

行業概覽

- 詞元質量及生產效率決定有效供應能力，成為支撐AI應用規模化落地的關鍵

隨著大規模部署AI應用釋放更廣泛的詞元需求，行業競爭關鍵不再僅是底層算力資源規模，而在於能否將分散、異構的芯片資源、模型能力和部署環境高效轉化為穩定可用的詞元產出。詞元需求快速增長的同時，供應端的芯片、內存容量、電力、數據中心及算力集群建設等仍受到供應鏈、建設週期和資源配置等物理約束。在需求快速釋放而供應相對有限的雙向催化下，提升詞元生產效率、提高詞元服務的穩定性和可獲得性，成為支撐AI應用從試點探索走向規模化落地的關鍵能力。

AI應用爆發下的詞元需求變化演進



資料來源：弗若斯特沙利文

全球及中國詞元供應市場分析

詞元供應產業現狀及痛點

- 多元算力環境增加適配與編排複雜度，詞元質量及生產效率有待提升

從AI芯片體系來看，全球算力供應主要由英偉達及國產廠商等支撐。不同芯片在硬件架構、軟件生態、性能邊界、適配成熟度及成本結構方面存在顯著差異，一方面抬升同一模型跨芯片的部署適配、算子優化與性能調優門檻；另一方面跨芯片混合集群進一步加大全局算力資源編排難度，異構算力池難以均衡編排推理任務，極易出現硬件忙閒分化、資源閒置浪費。若企業不具備成熟的跨芯片適配、推理引擎深度優化與全鏈路工程調優能力，詞元吞吐量、集群資源利用率、效率、用戶體驗及單位詞元成本等核心指標都會出現大幅波動，這會導致底層算力無法充分釋放峰值性能，難以穩定承接海量AI的推理需求，進而推高詞元生產綜合成本，大幅削弱企業客戶規模化採購詞元服務的商業經濟性。由此，AI基礎設施層的競爭重點正從單純囤積算力資源，轉向比拼詞元生產效率。

行業概覽

- **模型與部署生態高度多樣化，需保證詞元服務的穩定性**

從模型體系來看，當前市場並存DeepSeek、Qwen、GLM、Kimi等多款AI模型，各模型在參數規模、架構設計、推理效率、上下文長度及部署要求上各有差異，應用場景上也有所區分。疊加企業客戶在部署方式、數據安全、模型選型上的多元化訴求，行業逐步形成跨芯片、多模型、多部署方式共存的推理基礎設施格局。對於企業客戶而言，詞元服務的實際價值不僅取決於模型能力本身，也取決於詞元能否被穩定、連續、低延遲地調用，並在不同業務場景中保持穩定體驗。因此需要中間基礎設施層通過集中管理、統一部署和編排各類異構算力和AI模型，保證詞元的生產速度、運行穩定性與場景適配性。

詞元供應平台在詞元生產全生命週期中的賦能

在多元算力、多模型及多部署方式並存的行業環境下，針對詞元生產效率波動、資源利用不均及服務穩定性要求提升等行業痛點，詞元供應平台作為連接算力、模型與應用的系統軟件層，對詞元生產和交付全過程進行組織、編排和優化。

- **詞元生產全生命週期：涉及基礎資源接入、應用需求接入、模型選型、異構算力編排、推理執行、詞元封裝及結果交付**

詞元生產是由算力資源編排、模型部署和應用交付共同組成的連續服務流程。首先，詞元生產需要接入能源、數據、模型及算力集群、高速互聯、液冷系統等基礎設施，形成可供調用的底層資源基礎。隨後，用戶通過API、應用接口、智能體或各類應用系統發起詞元調用需求，平台根據請求內容、任務類型、模型能力、上下文長度、響應時延及服務等級要求，完成模型選型、任務路由、資源匹配等。

進入推理環節後，平台進一步將任務分配至適配的異構算力資源、推理實例和運行環境，由推理引擎調用底層芯片、服務器及模型執行環境，並基於用戶請求生成輸出詞元。最後，生成結果被封裝為可計量、可定價、可交付的標準化詞元服務，並通過各種AI應用交付到用戶側，同時平台持續監控服務質量、資源利用效率及單位詞元成本，反向優化模型部署、算力編排和運營策略。

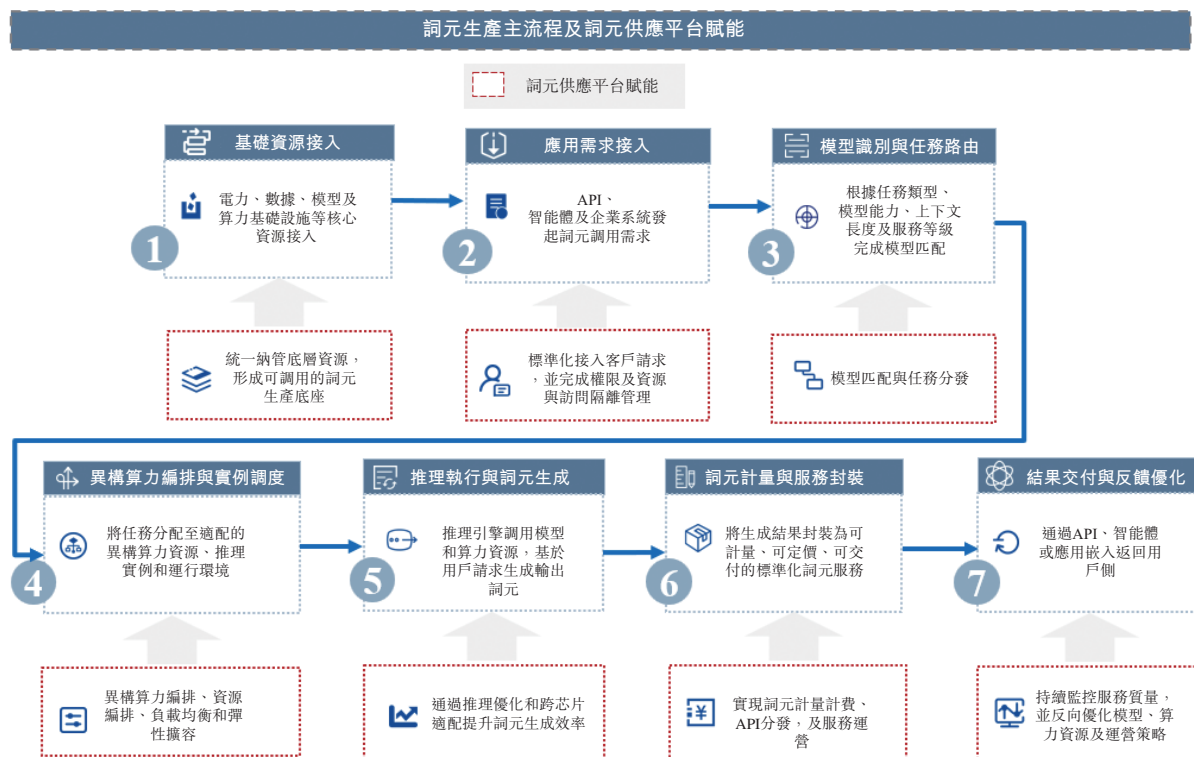
- **詞元供應平台賦能詞元生產全生命週期：具備提升詞元供應效率、穩定性與可擴展性的重要戰略價值**

在詞元生產鏈路中，詞元供應平台承擔連接底層算力資源、模型和應用的中樞作用，通過跨芯片適配、模型部署、推理優化、異構算力資源編排、彈性擴縮容、服務分發、計量計費及服務監控等能力，對詞元供應平台生產和交付進行全流程賦能，將複雜推理需求轉化為標準化、可管理、可規模化交付的詞元服務，從而提升詞元供應效率、降低跨芯片和多模型適配複雜度，並增強服務響應的穩定性和連續性。

行業概覽

對於客戶而言，詞元供應平台提供的不只是底層算力資源或單一模型調用能力，而是具備成本可控、穩定響應和多場景適配能力的封裝好、易用的詞元供應服務；對於算力和模型生態而言，平台則通過工程化優化和服務化交付，使底層異構算力和多模型能力更高效地轉化為可被上層應用持續調用的詞元生產能力。因此，詞元供應平台成為支撐AI應用規模化落地的關鍵基礎設施環節。

詞元生產主流程及詞元供應平台賦能示意圖



資料來源：弗若斯特沙利文

市場驅動因素分析

- 需求：差異化模型能力催生多元調用需求**

不同模型在領域專長、能力與適用場景上存在顯著差異，單一模型難以滿足全部業務需求。同一企業、同一業務線甚至單一產品，往往需要同時調用多款模型，以匹配不同任務的能力要求與適配場景。企業多元化的模型調用需求也推動詞元調用規模持續增長，催生對高適配性、多模型詞元服務的需求。

- 供應：降本增效需求推動專業化詞元供應服務發展**

隨著企業級AI應用進入生產環境，客戶對詞元獲取成本、響應速度、服務穩定性及調用效率的要求不斷提高。專業化詞元供應平台可通過推理優化、資源編排、緩存機制和服務運營等能力，提高詞元產出效率，降低詞元生成及調用成本，幫助客戶更高效地獲取AI能力。

行業概覽

• 技術：異構算力資源與多模型生態並存推動適配需求提升

市場同時存在英偉達及多款國產AI芯片等異構算力，以及DeepSeek、Qwen、Kimi、GLM、Wan等多元模型生態。產業政策導向、國產AI芯片發展、多元技術路線競爭及客戶部署需求分化共同塑造了長期多元異構格局，使跨芯片適配、模型部署、推理優化和編排能力的重要性進一步提升。

• 生態：企業級部署需求分化推動獨立生態價值提升

企業客戶對公有／本地部署／運營、資源與訪問隔離以及場景化適配的需求日益多樣化，推動詞元供應服務由單一服務模式向多模式發展。獨立生態詞元供應平台能夠連接多類型算力、多元模型及差異化企業客戶部署需求，在多方競合格局中提供更靈活的模型選擇、部署方式和詞元服務能力。

行業主流服務模式

詞元供應平台的服務模式通常與客戶調用規模、資源佔用方式、服務性能要求及部署方式密切相關。根據服務形式差異，行業內主要形成無服務器詞元服務、專屬實例及本地部署解決方案三類服務模式。

在該等服務模式中，無服務器詞元服務以共享資源的方式提供詞元服務，一般按所消耗詞元的使用量計費，主要面向尋求靈活且具成本效益的AI能力接入途徑的開發者、中小型企业及其他客戶。專屬實例在公有雲基礎設施上提供預留或專屬的算力，定價基於預留算力及底層基礎設施成本，旨在滿足對性能、穩定性有較高要求的客戶。本地部署解決方案將我們的推理引擎及算力資源調度系統部署於客戶自身的數據中心或專屬算力環境中，使客戶能夠在其自有算力資源上使用廣泛模型的詞元服務，主要面向大型企業及機構客戶。

行業主流服務模式

	無服務器詞元服務	專屬實例	本地部署
服務模式	基於共享資源提供標準化詞元服務，客戶通過單獨標準接口接入，無需預先投入專屬資源	為特定客戶於公有雲基礎設施上預留或專屬算力	將推理引擎及算力編排系統部署於客戶自有的數據中心或私有算力環境內
計費基礎	使用量按所消耗的詞元計量；通常為預付餘額，可為若干經常客戶提供後付按月結算	預留算力，當中考慮到底層基礎設施成本，包括算力及相關營運開支	軟件許可費及／或專業實施或維護服務費，視乎部署規模、系統配置及客戶需求而定
服務特徵	靈活且具成本效益的接入方式，適用於測試、驗證及彈性調用場景	較高的性能、穩定性，適用於對延遲敏感及高吞吐量的AI工作負載	採用標準化及模塊化的平台架構（僅提供有限的定制），適用於特定的部署環境或營運要求
適用客戶	尋求以靈活且具成本效益的方式獲取AI能力的開發者、中小型企业及其他客戶	運行對延遲靈敏、高吞吐量的AI工作負載的企業用戶，以及對性能和穩定性有更高要求的其他客戶	大型企業客戶及機構客戶

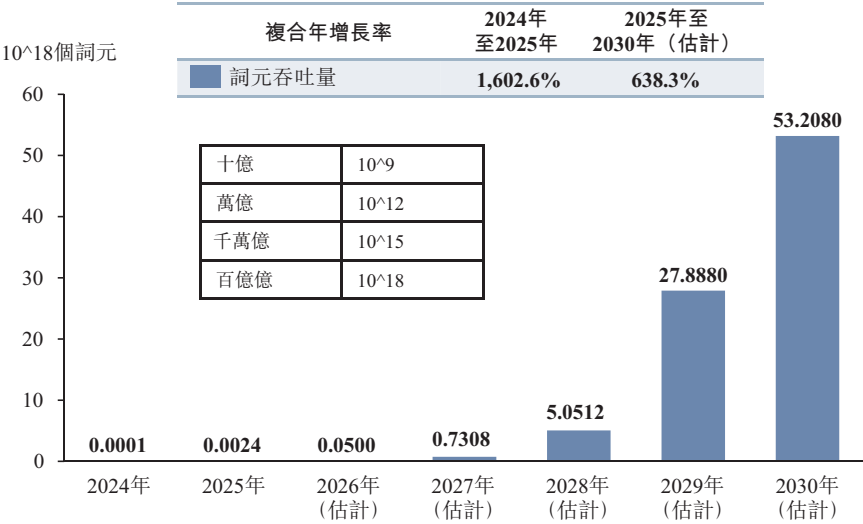
資料來源：弗若斯特沙利文

行業概覽

詞元供應市場規模

2024年，中國詞元吞吐量約為142.5萬億次，2025年快速提升至約2,426.3萬億次，2024年至2025年增長率達到1,602.6%。這一增長主要源於AI模型能力持續提升後，AI使用形態由單次問答和內容生成，逐步向連續任務執行、複雜目標拆解、多步驟工具調用及多模態交互延伸，推理調用頻次、上下文長度、併發需求和任務複雜度同步提升。進入2026年後，詞元消耗量預計將進一步躍升至約50千萬億次水平；預計在2030年，中國詞元消耗量將達到約53.2百億億個，2025年至2030年複合增長率約為638.3%。在此背景下，詞元供應市場將釋放長期、高確定性的市場增長潛力。

中國詞元供應市場規模（按詞元調用量，單位為百億億），2024年至2030年（估計）



資料來源：弗若斯特沙利文

市場發展趨勢分析

• 算力供應：算力持續呈現多樣化

未來的算力供應將維持多樣化的格局，其特徵為國外與國產AI芯片並存，以及存在多個供貨商及多種並行硬件架構（如GPU及NPU）。隨著算力異構性及生態系統適配持續發展，多源、多供貨商及多架構的異構算力資源編排能力將成為行業的核心焦點。

• 服務模式：詞元供應服務將由單一無服務器詞元服務向多模式並行發展

隨著企業客戶從測試調用走向生產級部署，其對調用規模、響應穩定性、延遲容忍度及資源與訪問隔離能力的要求不斷分化。詞元供應平台的服務模式將從共享資源池下的詞元吞吐量，逐步擴展至詞元吞吐量、專屬實例及本地部署並行，以滿足彈性伸縮、穩定供應和私有部署等不同需求。

行業概覽

- **產品能力：詞元生產能力將圍繞推理效率和服務穩定性持續深化**

在多模型、多芯片、多部署方式並存的環境下，詞元供應平台需要圍繞詞元生產全流程提升技術能力。推理引擎層將提升詞元產出效率，集群編排層將優化算力資源利用率，MaaS及API分發／運營層將強化多模型接入、統一調用、服務監控及客戶運營能力，從而提升詞元服務的效率、穩定性和可擴展性。

- **生態：獨立生態平台將在多元異構格局下承擔更重要的連接作用**

雲服務廠商、大型科技企業及模型廠商多依托自有資源構建封閉生態，服務體系優先适配自身算力、模型及應用；獨立生態詞元供應平台則秉持中立定位，不向單一算力或模型傾斜資源，而是連接多類型算力、AI模型及產業鏈各方主體，在跨芯片、多模型、多部署環境的服務交付中，形成獨特的價值。

企業關鍵成功因素分析

- **跨芯片、多模型适配及推理優化能力**

在多類型算力與多模型並存的市場環境下，企業需要具備對不同芯片架構、軟件生態、推理框架及模型結構的理解和适配能力。通過跨芯片适配、算子優化、模型部署優化及推理引擎優化等方式，企業能夠充分挖掘算力潛力，提升詞元產出效率，降低詞元生產成本，並增強詞元服務的可用性和穩定性。

- **集群編排、彈性服務及穩定交付能力**

隨著AI應用由單次調用向連續調用、複雜任務執行和生產級部署延伸，詞元服務需要在高併發、低延遲和成本可控的條件下穩定運行。企業需要具備兼容各類算力資源的編排能力，可統一納管並高效調度第三方算力中心、內部算力及本地部署算力。在此基礎上，依托GPU服務器及異構算力集群的統一管理、資源編排、實例調度、彈性伸縮、服務監控及運維管理能力，提升資源利用率並保障生產級服務交付。

- **中立開放與多元生態整合能力**

詞元供應平台需要連接芯片公司、算力集群供應方、模型公司、開發者及企業客戶，形成多元算力資源、多模型、多部署環境和多場景的詞元供應能力。平台聚焦基礎設施服務，與各類應用廠商、終端客戶不存在業務競爭關係。相較於封閉生態主體兼具應用開發與服務提供的雙重屬性，獨立生態詞元供應平台定位更為中立，能夠開放地聯動產業鏈各類參與主體開展深度協作。

行業概覽

中國詞元供應市場競爭分析

市場競爭格局概覽

中國詞元供應市場處於快速發展和競爭格局形成階段，行業競爭主要圍繞詞元生產效率、異構算力編排能力、多模型適配能力、企業級服務及連接能力展開。目前，市場參與主體可分為獨立生態與封閉生態兩類，具有資源稟賦、生態定位及服務邊界差異。

詞元供應市場不同生態對比

- 詞元供應市場主要形成獨立生態與封閉生態兩類商業模式

獨立生態詞元供應平台通常不綁定任何雲提供商、模型或應用場景，而是通過連接多類型算力、多元模型、不同算力中心及企業客戶需求，提供多模型、多算力、跨部署環境的詞元供應服務。封閉生態則通常由雲服務廠商、大型科技企業或模型廠商依托自有雲資源、模型、算力基礎及應用場景形成，能夠在自身生態內實現資源整合、模型調用和服務閉環。

- 獨立生態在多元異構格局下具備中立開放及跨生態連接價值

相較於封閉生態，獨立生態詞元供應平台具備中立開放、多元連接的特性和橫向適配能力，客戶可以根據需求自由組合算力與AI模型。這種中立定位在AI產業中早期的多方競合格局中更易獲得客戶信任，並能跨越不同生態實現連接，從而沉澱出封閉體系難以替代的差異化價值。

詞元供應市場參與主體對比						
特性對比	定位	資源基礎	應用關係	核心價值	兼容性	跨生態服務完整性
詞元供應企業	獨立生態	中立開放，連接多方生態	連接多類型算力，多元模型及不同數據中心	通常不與客戶或應用廠商競爭	跨模型、跨算力及跨場景的中立詞元供應	
	封閉生態	依託自有資源形成封閉生態	主要依託自有雲資源、模型體系、算力、及應用場景	自身可能擁有應用入口或行業解決方案，與部分外部應用廠商存在潛在業務邊界重疊	在自有生態內實現資源整合、模型調用和服務	

資料來源：弗若斯特沙利文

行業概覽

市場企業排名和市場份額分析

企業排名及市場份額

2025年，按年度詞元調用量計算，在中國詞元供應市場中，公司排名第四，市場份額約為1.5%。

中國詞元供應平台排名（以詞元調用量計），2025年

排名	公司名稱	詞元調用量 (萬億詞元)	市場份額 (%)
1.....	公司A	1,036.0	42.7
2.....	公司B	788.5	32.5
3.....	公司C	286.3	11.8
4.....	本公司	36.7	1.5
5.....	公司D	34.8	1.4
	其他	243.9	10.1
	總計	2,426.3	100.0

附註：

- (1) 公司A為一家私營企業，成立於2020年，總部設於中國北京，主營業務聚焦雲與AI服務，核心業務包括雲基礎設施、視頻與內容分發、大數據、人工智能及開發運維等。
- (2) 公司B為一家私營企業，成立於2009年，總部設於中國浙江，其隸屬於一家上市集團，並提供雲計算及AI服務，服務主要涵蓋計算、存儲、網絡、數據庫、大數據、AI模型服務及雲原生服務。
- (3) 公司C為一家私營企業，成立於2001年，總部設於中國北京，其隸屬於一家上市集團並經營AI雲服務，其服務主要涵蓋雲端基礎設施、AI開發平台、模型服務及行業AI解決方案。
- (4) 公司D為一家私營企業，成立於2019年，總部設於中國上海，專注於分佈式雲計算服務，核心業務包括智能計算、模型服務、邊緣計算及邊緣CDN。

獨立生態詞元供應平台排名及市場份額

2025年，按年度詞元調用量計算，在中國詞元供應市場中，公司排名獨立生態詞元供應平台第一，市場份額約為1.5%。

行業進入壁壘分析

• 跨芯片、跨模型的無縫整合與性能優化壁壘

詞元供應平台需要在多類型芯片、多元模型及不同推理引擎之間實現無縫整合，相關能力涉及模型部署、算子優化、推理引擎優化及多算力資源編排等多個環節。由於不同芯片的軟件生態、算子兼容性和模型結構差異較大，相關能力通常依賴長期工程實踐和持續迭代積累，新進入者較難在短期內形成可復用、可規模化的適配與優化能力。

行業概覽

- **大規模詞元服務運營壁壘**

詞元供應服務需要在高併發、低延遲和成本可控的條件下穩定運行，對企業的集群編排、資源編排、彈性伸縮、負載均衡、服務監控及運維管理能力提出較高要求。尤其在生產級應用場景中，客戶對詞元服務的穩定性、響應速度和連續交付能力要求較高，缺乏大規模運營經驗的企業較難滿足客戶長期調用需求。

- **生態協同與資源獲取壁壘**

詞元供應平台需要連接芯片廠商、算力供應方、模型公司、開發者及企業客戶，形成跨算力、跨AI模型和跨場景的服務能力。生態資源的積累不僅取決於接入數量，也取決於企業能否在多方合作中持續構建合作夥伴生態。新進入者通常需要較長時間建立合作關係、獲取資源及建立客戶信任。

- **商業化驗證與客戶忠誠度壁壘**

詞元供應平台的能力須於多種服務模式下持續進行驗證，包括無服務器詞元服務、專屬實例及本地部署。透過反覆的模型測試、芯片適配、部署遷移、推理優化、算力資源編排及服務運營，平台提供商累積可重複使用的工程專門知識及交付工具鏈。隨著企業客戶日益依賴該等平台進行生產級AI推理，穩定的服務表現、累積的整合經驗以及較高的轉換成本有助增強客戶忠誠度，從而使經驗豐富的平台提供商能夠建立可持續的競爭優勢。