



# 〔投資研究報告〕

2026/7/16

## 人工智慧價格戰： 中國開放權重模型重設企業成本曲線

### 目錄

- 一、價格錨下移：多模型路由進入生產環境
- 二、合規紅線：成本套利加入雙向政治選擇權
- 三、毛利池遷移：模型收入、雲端用量與軟體護城河
- 四、硬體分工：GPU 維持核心，ASIC 滲透依賴架構穩定
- 五、投資映射
- 六、驗證儀表板

## 每次成功任務成本取代模型榜單， 合規紅線決定實際滲透

2026/7/16 新光投顧

中國開放權重模型已進入西方企業的正式工作流程。DoorDash 將部分低階任務交由 Kimi 執行；Siemens 同時使用 DeepSeek、Z.ai 與美歐模型；Airbnb 將 Qwen 導入客服；Anysphere 的 Cursor 使用 Kimi；Coinbase 則以 GLM 5.2 與 Kimi 2.7 作為內部模型路由的低成本預設選項。

企業人工智慧的競爭軸線正由絕對性能轉向每次成功任務成本。大量企業任務只需跨過內部品質門檻，低價模型配合路由、快取、重試與高階模型回退，即可取得更低的整體成本。Vercel 的生產數據已呈現價格彈性：DeepSeek 單月詞元占比由不足 1% 升至 17%，支出占比仍約 1%；Anthropic 則以 32% 的詞元占比取得 65% 的支出，前沿模型仍控制高風險與高複雜度工作負載。

中國模型對美國模型商的破壞力首先體現在價格基準，直接西方收入份額居次。低價、開放權重與自行部署能力，使企業取得新的議價工具；OpenAI、Anthropic 與其他封閉模型商將被迫擴大中階降價、批次折扣、提示快取與長約優惠。最高性能、安全責任與企業支援仍可保留溢價。

地緣政治合規折價將限制中國模型的西方滲透率。美國國會已調查 Airbnb 與 Anysphere，模型供應商遭列入實體清單的政策選擇權仍在，北京亦可能限制海外取得先進模型。成本套利必須加計監管、供應中斷、模型行為與聲譽風險。

人工智慧毛利池將沿四條路徑重分配：模型層單位價格下移；超大型雲端供應商承接多模型代管與推論增量；企業軟體的價值集中於系統寫入權與交易執行；推論硬體維持總量成長，GPU 與 ASIC 依工作負載穩定度分工。

投資定調：降低對單一模型高毛利寡占的信心；提高多模型雲端、治理安全、核心交易系統與通用推論硬體的權重；客製化 ASIC 保留條件式上行。

## 目錄

- 一、 價格錨下移：多模型路由進入生產環境
- 二、 合規紅線：成本套利加入雙向政治選擇權
- 三、 毛利池遷移：模型收入、雲端用量與軟體護城河
- 四、 硬體分工：GPU 維持核心，ASIC 滲透依賴架構穩定
- 五、 投資映射
- 六、 驗證儀表板

## 一、 價格錨下移：多模型路由進入生產環境

模型價格差距已達足以改變企業架構的量級。

DeepSeek V4 Flash 在部分平台的每百萬詞元輸入與輸出價格約為 0.14 美元與 0.28 美元，較可比較的 Anthropic 產品低約 20 至 50 倍。DeepSeek 官方 V4 Pro 輸出價格為每百萬詞元人民幣 6 元，OpenAI GPT-5.6 Sol 與 Anthropic Opus 4.8 的輸出價格分別為 30 美元與 25 美元。模型能力、推理深度與可靠度不同，名目價差仍足以迫使採購部門重新分配工作負載。

企業採購開始採用三層模型結構：

| 工作負載       | 模型配置            | 核心採購標準            |
|------------|-----------------|-------------------|
| 高風險決策、複雜代理 | 美國前沿封閉模型        | 任務成功率、安全、可靠度、責任歸屬 |
| 高頻率、中等複雜度  | 中國開放模型與美國中階模型競價 | 每次成功任務成本、延遲、可替換性  |
| 低風險、批次處理   | 小模型、自建模型、專用模型   | 吞吐量、利用率、單位推論成本    |

DoorDash 的配置已呈現此架構：Kimi 處理低階工作，Anthropic 模型保留最困難的任務。Coinbase 透過內部模型閘道器，將 GLM 5.2 與 Kimi 2.7 設為低成本預設模型，同時保留工程師升級至前沿模型的權限。模型使用量持續攀升，總支出卻降至高點附近的一半。

每次成功任務成本的計算涵蓋模型費用、推理詞元、工具調用、檢索、快取、重試、人工覆核與延遲損失。低價模型若產生較高失敗率，名目價格優勢會被重試與人工成本侵蝕；品質通過企業內部評測後，路由器將迅速把工作負載轉往低成本供應商。

公開榜單對採購決策的影響下降。企業將依自身客服一次解決率、程式測試通過率、文件修正時間、代理完整成功率與合規事故率建立內部評測。模型領先幅度無法轉化為相應的價格倍數時，工作負載便會分流。

## 二、合規紅線：成本套利加入雙向政治選擇權

美國國會對 Airbnb 與 Anysphere 的調查，將地緣政治風險由模型供應端推進至企業採用端。

Airbnb 使用 Qwen 的理由是速度與成本，國會委員會仍要求公司說明供應商接觸、風險審查、資料安全與模型選擇。Anysphere 因 Cursor 使用 Kimi 接受相同調查。正式禁令尚未落地，調查、聽證與文件要求已提高企業法務、資安與董事會的採用門檻。

大型企業的模式採購需通過四道紅線：

1. 模型公司或關係企業遭列入實體清單、軍方名單或制裁名單；
2. 政府承包商、金融、醫療與關鍵基礎設施被禁止使用中國來源模型；
3. 底層程式碼、訓練資料、模型更新或輸出行為被認定具有國安風險；
4. 中國政府限制模型權重、API、更新版本或關鍵人才向海外供應。

DeepSeek 等百餘家中國企業據報已通過美國跨部門實體清單審查，正式公布因外交考量暫緩。這是一項尚未執行的政策選擇權。禁令一旦啟動，美國受監管企業將迅速停止採用，AWS 等雲端平台也可能下架相關模型。

模型行為風險進一步提高合規折價。Booz Allen 對四款中國模型與一款美國模型進行逾 2,800 次測試；三款中國模型在提示使用者為美國政府人員時產出更多安全漏洞，中國模型亦對北京敏感議題出現拒答與立場偏移。研究尚未證明漏洞為蓄意植入，CISO 難以忽略身分相關的輸出差異。

北京的反向限制構成另一端供應風險。中國主管機關已與 Alibaba、ByteDance 及 Z.ai 討論限制先進模型海外存取、延後公開權重或限定境內使用。建立在中國模型上的西方應用，可能同時面對華盛頓禁用與北京斷供。

西方市場將分成兩個價格區：

- 受監管工作負載支付可信模型、資料駐留與責任承擔的合規溢價；
  - 一般商業工作負載使用開放模型、多模型路由與自建推論取得成本套利。
- 中國模型的直接西方市占受政治限制，價格破壞仍可透過美國開放模型、企業議價與封閉模型降價持續傳導。

### 三、毛利池遷移：模型收入、雲端用量與軟體護城河

**模型供應商：高端溢價保留，中階價格下移**

Vercel 數據顯示，DeepSeek 取得 17%的詞元量，支出占比僅約 1%；Anthropic 以 32%的詞元量取得 65%的支出。市場同時存在低價模型大量滲透與前沿模型高額支出。

OpenAI 與 Anthropic 仍控制複雜程式設計、後台代理與高風險任務，定價權將集中於任務成功率、安全責任、可靠度與企業支援。中階模型面臨商品化，單位收入與毛利率承壓。

模型商的收入增速取決於詞元量能否跑贏價格下降。代理工作負載包含規劃、搜尋、工具調用、平行代理與重試，單一任務的詞元需求高於傳統聊天。價格下降可釋放更多企業應用，總收入仍有成長空間。

**超大型雲端供應商：推論增量與資本效率並存**

AWS Bedrock 已代管 DeepSeek、Kimi、Qwen、GLM 與 MiniMax。模型供應商更換時，企業仍需雲端運算、儲存、網路、資料庫、監控、身分管理與安全服務。模型層商品化提高雲端平台的控制力。

AWS、Azure 與 Google Cloud 可同時承接多模型市集、路由、資料治理與推論收入。雲端收入受惠於詞元量擴張，毛利率受到資料中心折舊、電力、晶片採購與價格競爭壓力。自研晶片、軟體最佳化與利用率決定最終股東回報。

#### 企業軟體：交易執行權取代資料所有權

持有企業資料只能提供初步優勢。Microsoft Copilot 與 Google Workspace 代理可透過連接器讀取第三方資料，並逐步跨越原有軟體介面。只提供搜尋、摘要、分析與報表的軟體，容易被水平代理繞過。

高防禦力軟體同時掌握四項資產：

- 具法律與營運效力的正式系統紀錄；
- 修改 ERP、CRM、財務、人資與採購資料的寫入權；
- 建立訂單、核准付款、調整庫存與更新合約的交易執行權；
- 身分驗證、權限控制、版本紀錄與稽核責任。

SAP、ServiceNow、Salesforce、Workday 等平台具備系統紀錄與轉換成本，AI 價值需由分析延伸至核心交易執行。代理若只能提出建議，客戶支付意願有限；代理能安全完成採購、付款、客服結案或供應鏈調度，軟體可依完成任務與業務成果定價。

薄封裝應用承受功能商品化與水平平台降維競爭。核心交易系統的風險集中於 Microsoft 與 Google 直接取得寫入權，以及企業將流程控制層上移至獨立代理平台。

## 四、硬體分工：GPU 維持核心，ASIC 滲透依賴架構穩定

低價模型提高詞元需求，也增加工作負載異質性。

模型架構持續從稠密 Transformer 轉向混合專家、稀疏運算、長上下文與代理協作。每一輪模型變化都會改變記憶體容量、頻寬、資料流與算子需求。硬體通用性在架構尚未收斂的階段具有保險價值。

NVIDIA 的護城河集中於 CUDA、編譯器、推論引擎、函式庫與模型相容性。相同 GPU 叢集可切換模型、精度、批次與工作負載，降低專用晶片因模型改版而閒置的風險。多模型路由普及反而提高通用運算平台的價值。

ASIC 的成本優勢需要三項條件：

1. 工作負載量體龐大且長期穩定；
2. 模型架構與運算圖具備足夠生命週期；
3. 晶片利用率足以攤銷設計、流片、軟體移植與維護成本。

搜尋、推薦、廣告、固定版本模型與雲商內部服務適合客製化晶片。快速迭代的前沿模型、企業長尾模型與跨模型路由仍偏向 GPU。架構改版可能削弱 ASIC 利用率，份額替代將呈現漸進路徑。

Broadcom 與雲商自研晶片供應鏈保有長期成長，近期估值需要量產時程、客戶集中度、利用率與次世代延續性驗證。NVIDIA 仍是推論總量擴張的核心表達，HBM、網路交換、光通訊、電力與散熱承接 GPU 與 ASIC 共同增加的基礎設施需求。

## 五、投資映射

| 產業環節                   | 定價方向        | 核心條件                     |
|------------------------|-------------|--------------------------|
| 美國前沿模型商                | 收入續增、單位毛利承壓 | 高端任務與合規市場保留溢價，中階價格競爭加劇   |
| 中國開放模型商                | 用量成長、海外變現受限 | 價格錨有效，直接市占受美中雙向政策限制      |
| AWS、AZURE、GOOGLE CLOUD | 收入正面、毛利分化   | 多模型代管與推論量成長，自研晶片與利用率決定回報 |
| 資安、評測與治理平台             | 正面          | 模型原產地、行為測試、權限與稽核需求上升     |
| 核心交易型 SAAS             | 選股正面        | 寫入權、執行權、轉換成本與責任承擔        |
| 通用生產力與薄封裝 SAAS         | 偏負面         | 水平代理繞過功能層，模型成本下降壓縮定價     |

|                  |       |                 |
|------------------|-------|-----------------|
| NVIDIA GPU       | 中期正面  | 多模型與架構變動提高通用性價值 |
| 客製化 ASIC         | 條件式正面 | 固定、高量、長生命週期工作負載 |
| HBM、網路、光通訊、電力與散熱 | 正面    | 總推論量與代理複雜度持續上升  |

Microsoft 同時具備 Azure 推論受益、OpenAI 經濟曝險與 Copilot 對 SaaS 的侵蝕能力，方向最為混合。Amazon 與 Alphabet 在模型中立代管及自研晶片上具備較高純度。NVIDIA 與 Broadcom 分別代表通用推論與專用推論，短期勝負取決於模型架構變化速度。

## 六、驗證儀表板

| 觀察變數               | 商品化加速            | 寡占或合規溢價維持          | 投資含義        |
|--------------------|------------------|--------------------|-------------|
| DEEPSEEK 詞元 / 支出占比 | 詞元份額續升、支出占比維持低檔  | 高風險任務仍集中 Anthropic | 量增與收入分離     |
| 中美模型價格差            | 維持 10 倍以上        | 中階價差快速收斂           | 模型層毛利壓力     |
| 企業路由採用             | Coinbase 式自動路由普及 | 單一模型長約維持           | 雲端與治理層定價權   |
| 採用端國會調查            | 調查擴及更多企業         | 調查降溫、無新增限制         | 合規寒蟬效應      |
| DEEPSEEK 實體清單      | 正式列名或政府禁用擴大      | 公布持續延後             | 中國模型西方滲透上限  |
| 中國海外供應政策           | 限制權重、API 或版本更新   | 維持全球開放             | 反向供應鏈中斷風險   |
| 模型行為安全測試           | 身分相關漏洞重複出現       | 第三方測試無系統差異         | CISO 否決機率   |
| SAAS 交易執行率         | AI 直接寫入與完成交易     | 功能停留於摘要分析          | 軟體護城河強弱     |
| GPU / ASIC 推論份額    | 固定工作負載轉向 ASIC    | 新模型持續依賴 GPU        | 專用晶片替代速度    |
| 雲端 AI 收入與毛利率       | 用量增速超過價格下降       | CapEx 與折舊壓低毛利      | 收入成長能否轉化為獲利 |

## 結論

中國開放權重模型已重設企業人工智慧的價格基準。DeepSeek 在 Vercel 取得 17% 的詞元量、支出占比僅約 1%，證明低價模型已跨過生產環境的品質門檻。前沿模型仍控制高風險任務與大部分支出，模型市場將形成高端溢價與中階商品化並存的結構。

地緣政治決定實際滲透率。國會對 Airbnb 與 Anyosphere 的調查，顯示政策壓力已打到採用端；Booz Allen 的程式碼安全測試提高 CISO 否決機率；美國實體清單與中國海外供應限制構成雙向政策選擇權。中國模型可取得的西方收入受到壓制，其全球價格破壞力仍將延續。

軟體毛利池向掌握系統寫入、交易執行與稽核責任的平台集中。專有資料缺乏獨立防禦力，水平代理可透過連接器繞過分析型介面。核心營運交易與轉換成本決定長期定價權。

推論基礎設施維持總量多頭。GPU 通用性在模型架構快速變動期具備較高價值；ASIC 在固定、高量與長生命週期工作負載中逐步滲透。模型價格下降將推升詞元與代理用量，雲端、HBM、網路、電力與散熱承接毛利池下移。

產業主線已由模型能力排名，轉向成本、合規、工作流程控制權與硬體利用率。模型供應商的高毛利寡占開始鬆動，人工智慧總需求仍在擴張。

**點我加入新光證券官方 Line 帳號**，每週第一時間收到新光投顧免費總經、產經報告