

技術詞彙表

本詞彙表包含本文件中與本公司及其業務相關的若干技術術語的定義。這些定義可能與行業標準定義不符，也可能與其他公司採用的類似術語不具可比性。

「AGI」	指 通用人工智能，一種理論上的人工智能形式，具備以等於或超越人類水平，跨越廣泛任務領域理解、學習和應用知識的能力
「AI」	指 人工智能，由機器（特別是計算機系統）對人類智能過程的模擬
「AI智能體」 或「智能體」	指 能夠自主理解、規劃和執行複雜工作流程或任務的人工智能系統，從簡單的文本生成過渡至自主執行
「AI應用」	指 利用人工智能技術執行特定任務或提供服務的軟件應用程序
「AI能力」	指 由人工智能系統所提供的智能屬性或智能能力
「AI基礎設施」	指 開發、訓練和部署AI模型所需的底層硬件和軟件資源，包括算力、存儲和網絡
「AI模型」	指 基於數據訓練的數學或計算模型，用於識別模式並做出預測或決策
「AI原生」	指 指從底層設計上就以人工智能為核心組成部分而構建的產品、服務或基礎設施，而非將其作為附加功能
「AIGC」	指 人工智能生成內容，利用AI算法生成文本、圖像、音頻或代碼等新內容
「API」	指 應用程序編程接口，一套允許不同軟件應用程序相互通信和交互的規則與協議
「API調用」	指 客戶端應用程式向應用程序編程接口提交請求並接收回應的過程，通常供客戶用於存取及使用我們平台上的AI模型推理服務
「ASIC」	指 專用集成電路，為特定用途定制設計，而非用於通用目的的集成電路芯片
「CAGR」	指 複合年增長率

技術詞彙表

「封閉生態」	指 由雲服務提供商、大型科技公司或模型企業依托其自有雲資源、AI模型、算力資源及應用而建立的封閉生態AI詞元供應模式。於該模式下，詞元供應服務、模型調用、算力資源分配及應用整合主要於服務提供商自身生態系統內完成，形成內部閉環
「計算圖」	指 表示機器學習模型中數學運算和數據流的有向無環圖
「算力資源」	指 硬件提供的用於執行計算任務的處理能力和內存
「算力資源編排」	指 對各種算力進行自動化安排、協調及管理，以提供集成化服務
「上下文」	指 提供給AI模型的背景信息和先前的對話歷史，用於生成相關且連貫的響應
「上下文長度」	指 大型語言模型在單次交互中能夠處理和保留的最大詞元數量，包括輸入提示和生成輸出
「跨芯片」	指 在一個系統或架構中涉及或集成多種類型的芯片
「跨芯片適配」	指 修改或配置軟件和模型，使其能在不同類型或多個芯片上高效運行的過程
「解碼」	指 AI模型推理過程中的解碼階段，模型基於已填充的上下文逐個生成輸出詞元
「專屬實例」	指 一種雲計算服務方案，特定的計算容量和硬件專為某一客戶獨佔使用，不與其他用戶共享，以確保數據隔離和性能穩定
「稠密模型」	指 一種傳統神經網絡架構，在處理每個詞元或輸入時，所有參數和計算節點均被激活，與稀疏模型相對
「部署」	指 安裝、配置並啟用軟件或模型使其在實際環境中運行的過程
「國產芯片」	指 在中國設計的半導體芯片
「雙碳」	指 中國實現2030年前碳達峰、2060年前碳中和的國家氣候目標
「動態量化」	指 一種動態降低在推理過程中神經網絡權重和激活值數值精度的技術，可在不顯著損失精度的情況下降低計算和內存需求

技術詞彙表

「彈性伸縮」	指 系統根據不斷變化的工作負載自動增加或減少算力資源的能力
「彈性服務」	指 能夠根據當前需求自動擴大或縮小算力資源規模的服務
「企業級」	指 指在可擴展性、可靠性和安全性方面旨在滿足大型組織嚴格要求的產品或服務
「專家並行」或「EP」	指 一種用於AI模型的分佈式計算技術，將不同的專用子網絡（即「專家」）分佈於多個計算設備上，以提高計算效率
「GPU」	指 圖形處理器，一種專為快速處理和變換內存以加速圖像創建而設計的專用電子電路，廣泛用於處理AI模型訓練和推理的大規模並行計算工作負載
「硬件架構」	指 計算機硬件組件及其互連的邏輯結構和物理設計
「硬件利用率」	指 硬件資源被積極用於有效工作的時間或容量百分比
「異構」	指 由不同或多樣的元素組成；在計算中，指使用不同類型處理器或架構的系統
「異構算力」	指 在單一算力環境中使用來自不同廠商的多種不同類型處理器或核心（如GPU、NPU和ASIC），以優化性能和成本
「異構算力資源」	指 來源於不同類型硬件架構或處理器的算力資源
「高速互聯」	指 實現不同硬件組件或節點之間快速數據傳輸和通信的技術或基礎設施
「IDC」	指 互聯網數據中心，用於存放計算機系統及相關組件（如通信和存儲系統）的物理設施
「獨立生態」	指 一種獨立生態AI詞元供應平台模型，不受限於任何單一雲服務提供商、算力資源、AI模型或應用。該模型以中立方式連接異構算力資源、多元AI模型及不同部署環境，並憑藉其橫向適配能力，提供跨不同生態系統的詞元供應服務
「推理」	指 使用已訓練的AI模型，根據用戶的新輸入提示詞進行預測、解決任務或生成輸出的過程

技術詞彙表

「推理引擎」	指 設計用於執行AI模型並基於輸入數據產生輸出（推理）的軟件系統
「推理優化」	指 用於在保持準確性的同時提高模型推理速度和效率的技術和過程
「KV緩存」	指 鍵值緩存，AI模型推理過程中使用的一種內存優化技術，通過存儲已計算的注意力鍵值對，避免多輪交互中的重複計算
「大規模部署」	指 在大量節點或環境中安裝和運行系統或模型
「延遲」	指 用戶向AI模型提交請求，到系統提供相應響應之間的時間差
「液冷」	指 使用液態冷卻劑從計算硬件中散熱的冷卻技術，可提高能源效率和密度
「負載均衡」	指 將工作負載分配到多個算力資源的過程，以確保最佳資源利用率並防止過載
「MaaS」或「模型即服務」	指 一種雲計算服務模式，通過API或平台向用戶提供對預訓練AI模型的訪問
「混合專家模型」或「MoE」	指 一種深度學習架構，使用多個被稱為「專家」的專用子網絡解決問題，對任何給定輸入僅激活部分專家，以提高計算效率
「模型能力」	指 AI模型能夠執行的特定技能和功能，如自然語言理解、圖像識別或推理
「模型部署」	指 將訓練好的AI模型集成到現有生產環境中以進行預測或決策的過程
「模型權重」	指 AI模型內部通過學習獲得的參數，決定了輸入數據如何轉化為輸出預測
「多模型」	指 集成或使用多種類型的AI模型
「多模態」	指 指能夠處理、理解和生成多種不同類型數據輸入和輸出（如文本、圖像、音頻和視頻）的AI模型或系統
「NPU」	指 神經網絡處理器，專為加速機器學習算法和神經網絡運算而設計的專用微處理器
「可觀測性」	指 基於日誌、指標和鏈路追蹤等外部輸出，測量和理解系統內部狀態的能力

技術詞彙表

「本地部署」	指 一種軟件交付模式，平台和AI模型完全安裝並運行於客戶的私有數據中心或本地基礎設施內
「算子」	指 神經網絡算法中的一種基礎數學函數或計算例程，對其進行深度優化，對於使AI模型能夠在特定芯片架構上高效運行至關重要
「編排」	指 對複雜計算機系統、中間件和服務進行自動化配置、協調和管理
「流水線並行」	指 一種分佈式計算技術，將AI模型按順序層級劃分到多個處理器上，允許不同微批次數據同時並行處理
「PoC」或「概念驗證」	指 為實現某種方法或想法以證明其可行性，通常用於驗證概念或理論在現實應用中的潛力
「預填充」	指 AI模型推理階段，模型處理輸入提示詞並預填充KV緩存以生成第一個詞元
「預填充－解碼分離」 或「PD分離」	指 一種推理優化技術，將用戶提示詞的初始處理（預填充）與後續輸出詞元的生成（解碼）分離至不同計算節點，以最大化集群級調度效率
「生產級」	指 足夠健壯、可靠且安全，能夠部署在實際終端用戶環境中的系統或軟件
「提示詞」	指 用戶向AI模型提供的輸入文本、指令或數據，用於引導模型生成響應或執行任務
「公有雲服務」	指 由我們通過互聯網向多名用戶提供的雲計算服務
「資源調度」	指 隨著時間推移將可用資源分配給任務的策略和機制，以優化整體系統性能
「無服務器」	指 一種雲計算執行模式，雲提供商按需動態分配算力資源，使客戶無需管理底層基礎設施即可構建和運行AI應用
「SLA」	指 服務級別協議，服務提供商與客戶之間的承諾，規定預期服務水平，通常包括可用性、正常運行時間和性能穩定性等指標
「系統軟件」	指 旨在運行計算機硬件並為應用程序提供和維護運行平台的計算機程序

技術詞彙表

「吞吐量」	指 計算系統或AI模型在給定時間內處理和生成的數據量或詞元數量
「詞元」	指 AI模型在推理過程中處理的基本數據單元，類似於自然語言中的單詞或子詞，既是生成的輸出，也是衡量計算消耗和計費的基本指標
「詞元消耗」	指 在特定操作或會話期間，AI模型處理或生成的詞元數量
「詞元供應平台」	指 一種系統軟件平台，通過組織、編排和優化整個詞元生成與交付流程，連接硬件資源、模型能力和應用需求，以提供穩定、高效、可擴展的詞元服務
「詞元/秒」	指 每秒詞元數，衡量AI模型推理生成速度和吞吐量的標準指標
「訓練」	指 使用數據集教授AI模型識別模式並做出準確預測的過程