即時發佈

阿里巴巴推出 Qwen4-Max

打造全新 AI 工作、創意及探索夥伴

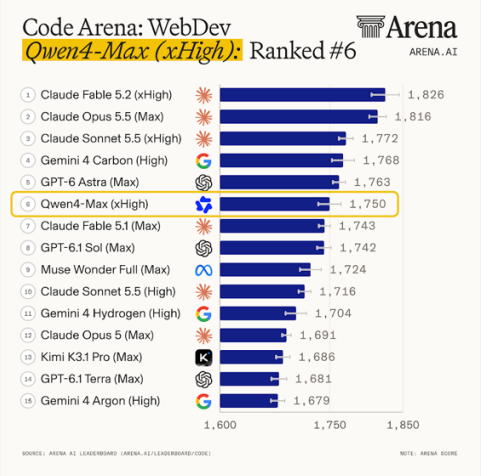
香港，2026 年 11 月 17 日 – 阿里巴巴集團正式發佈 Qwen4-Max，這是千問系列至今性能最強的模型。新模型總參數達 2.4 萬億，具備原生視覺能力，支援長達 100 萬個 token 的上下文視窗，擅長處理工作、創意及研究領域的複雜任務。

Qwen4-Max 在多項主流基準測試中名列前茅，在編程能力、智能體任務及視覺推理方面尤其突出。新模型在編程能力 Code Arena 榜單位列第六，並在 Agents' Last Exam、Vals Index、JobBench 及 Chartography 等涵蓋智能體工作流程、職場生產力及視覺推理的專項測試中取得佳績。

阿里巴巴的 Work360 基準測試涵蓋 721 個細分領域，評估模型處理複雜行業任務的表現。結果顯示，Qwen4-Max 的整體表現優於上一代 Qwen3.8-Max。在高強度推理模式下，token 用量減少 35%，每項任務的平均成本下降約六成。

這些提升源於全新的 Qwen4 架構，其稀疏混合專家（MoE）設計在注意力機制、殘差連接及嵌入層上均有改進。**Qwen4-Max 的總參數與上一代同為 2.4 萬億，每個 token 僅激活 710 億個參數，較 Qwen3.8-Max 少 25%，以更低的推理成本提供更強的性能。**

Qwen4-Max 現已在 QwenCloud 上線，用戶可透過 Token Plan 或 API 使用，亦可在 QwenWork 及 QwenStudio 等阿里巴巴平台體驗新模型。輸入（隱式緩存命中）每 100 萬個 token 收費 0.075 美元。**Qwen4 系列中規模較小的開放權重模型亦將於短期內向全球開發者發佈。**



Qwen4-Max 在 Code Arena 排行榜位列第六



Qwen4-Max 在 Work360 的整體表現優於 Qwen3.8-Max

提升資源效率及生產力

Qwen4-Max 具備先進的雲端資源調度能力，可部署於雲端基礎設施。模型以閉環方式自主優化自身的調度流程，降低成本並提高資源使用效率。基準測試顯示，該方案在數據標註場景中令模型成本費用平均下降 9%，將 CPU 離線重調度場景的**最優求解速度**提升逾 14 倍。

除應用於雲端基礎設施外，**Qwen4-Max 亦支援專業工作與創意製作**，涵蓋 3D 建模、遊戲開發及透過智能體製作影片等場景。在**汽車製造**方面，模型可支援生產線規劃、零件設計以至加工驗證的全流程。在一次演示中，模型根據回饋反覆調整加工方案，在正式加工前找出切削過量及材料去除不足等問題，將單個工件加工編程的人工用時由約 475 分鐘縮短至 24 分鐘，減幅達 95%。Qwen4-Max 亦能加快金融、法律及醫療等行業的前線部署工程（FDE）。

支援智能體持續運作 推動科學發現

Qwen4-Max 為智能體時代而設，支援全天候持續執行任務。模型能規劃長週期任務，有助科研團隊銜接數碼探索與實體實驗。為應對新一代晶片的散熱挑戰，**Qwen4 智能體連續運作 48 小時，完成四輪閉環流程，**結合分子設計、合成與測試及主動學習，**成功找出一種經實驗驗證的冷卻液分子。在一項阿茲海默症研究實驗中，Qwen4-Max 展示了**從靶點研究、分子設計到運算評估的**全流程能力**。模型在過程中自主調用工具 454 次，完成 3,239 項指標評估，用以設計和優化 24 個候選分子。

將安全融入模型設計

阿里巴巴將模型管治貫徹至模型的整個生命週期。**在軟件工程訓練中，功能測試與程式碼安全檢查同步進行，**大幅減少涉及注入攻擊、記憶體安全及訪問控制失效的漏洞。**網絡安全方面，阿里巴巴按任務目的篩選訓練數據，保留用於偵測及修補漏洞的防禦資料，並移除攻擊性內容。阿里巴巴亦採用專門的安全數據集，並在後訓練階段加強防護，進一步降低化學、生物、放射及核（CBRN）相關風險。**

###

關於阿里巴巴集團

阿里巴巴集團是一家聚焦 AI+雲及商業的全球科技公司。我們依托全棧 AI 能力與服務，從應用到算力基礎設施，為消費者及企業客戶提供支持。阿里巴巴 AI 科技以千問大模型及多模態模型系列為基石，為企業級解決方案及消費者平台等各項服務提供智

能驅動力。我們的電商業務堅持用戶為先，為商家、品牌、零售商及中小企業提供技術及營銷觸達能力，助力其連接客戶、高效運營。