

{% include header.html %}

{% include ai-share.html %}

{% include price.html %}

LLM

大語言模型 (Large Language Model)

- [AlpacaEval Leaderboard](#)
 - [Open LLM Leaderboard](#)
 - [Big Code Models Leaderboard](#)
-

- [Awesome-Chinese-LLM](#)
-

- [大模型所需GPU 記憶體筆記](#)
 - [不同參數規模大語言模型在不同微調方法下所需的顯存總結](#)
-

- [人工智慧大語言模型微調技術：SFT 監督微調、LoRA 微調方法、P-tuning v2 微調方法、Freeze 監督微調方法](#)
 - [LoRA、完全微調到底有何不同？MIT 21頁論文講懂了](#)
 - [大模型微調 \(Fine-tuning \) 全解 · 需要了解的都在這裡](#)
 - [初學者必看大模型微調指南：Unsloth官方微調技巧大公開！](#)
 - [零代碼！一站式完整資料集準備到模型微調全流程！](#)
 - [把你的DeePseek-R1 微調為某個領域的專家？](#)
 - [使用 NVIDIA NeMo 框架進行 LLM 模型剪枝和知識蒸餾](#)
 - [LLaMA Factory](#)： <https://huggingface.co/spaces/hiyouga/LLaMA-Board>
 - [官方 README_zh.md](#)
 - [2024-09-13：解析Llama-Factory：從微調到推理的架構](#)
 - [單卡3 小時訓練專屬大模型Agent：基於LLaMA Factory 實戰](#)
 - [torch tune](#)： <https://github.com/pytorch/torch tune>
 - <https://pytorch.dev.org/torch tune/stable/index.html>
 - [使用知識蒸餾將Llama3.1 8B 蒸餾到Llama3.2 1B](#)
-

- [微調特定領域的大模型 · 資料集究竟要怎麼搞？](#)
 - [LLaMA Factory 微調教學：如何建立高品質資料集](#)
 - [大模型微調資料集生產工具 Easy Dataset](#)
 - [開源DeepWiki版支援根據現有檔案產生微調資料集](#)
 - [零一萬物發布COIG-CQIA：高品質且符合人類互動行為的中文指令微調數據](#)
-

- [AI Search Has A Citation Problem](#)
- [AI Agents vs. Agentic AI: A Conceptual Taxonomy, Applications and Challenges](#)
- [OWASP Agentic AI – Threats and Mitigations](#)

Tool 名稱	功能範疇	集成能力	適用場景	建議選擇	知識庫技術
Flowise	簡單可視化流程建構	支持 LangChain 和 GPT，多平台部署	快速構建簡單 LLM 流程應用	適合需要快速構建和部署 LLM 應用的用戶	
Langflow	多智能體與 RAG 應用構建	外部工具、API 與資料庫集成	複雜任務鏈與多智能體應用開發	適合構建複雜任務鏈的技術型開發者	
Dify	全面應用開發與管理	多模型支持，全面工作流與模型管理	全生命周期管理與複雜應用開發	適合需要全面管理 LLM 應用的開發者或企業	基於 Pinecone 的向量數據庫、Notion API 同步（增量更新）、支持 Rerank 模型（bge-reranker-base）、提供行業模板庫
n8n	通用自動化與流程編排平台	支持超過 350 種服務與 API 集成，可視化流程編輯	應用整合、自動化工作流程構建、自動回應觸發器	適合需要自動處理非 AI 任務或整合各類 SaaS 工具的用戶與開發者	可與向量資料庫結合使用，但非內建
RAGFlow	模組化 RAG 管線與知識應用框架	支持 LangChain、Chroma、FAISS 等，可用於構建完整 RAG 工作流	知識問答、文件檢索、RAG 多階段優化	適合需要建構可定製、模組化 RAG 系統的開發者	支持多向量資料庫（Chroma、FAISS）、可結合自定義資料源與檢索策略
New API	模型接口統一與分發	OpenAI 格式統一，支持多支付協議與分發管理	多模型接口管理與分發	適合需要統一管理多種 AI 模型接口的用戶	

Tool 名稱	功能 範疇	集成能力	適用場景	建議選擇	知識庫技術
XORBITS Inference	分散式推理與部署	與 Hugging Face 等模型相容，支援雲端及本地等多種部署環境	大規模模型推理與雲端部署，需快速搭建可擴展的推理服務時	適合需要高效擴展能力、進行大規模模型推理的團隊或企業	
Ollama	本地模型推理與管理	提供命令列介面，支援多種 Llama 模型於本地運行	在有隱私或離線需求的場景下進行本地推理	適合想在本地快速配置 Llama 系列模型的個人或中小型團隊	
FastGPT	知識庫問答系統與工作流編排	提供開箱即用的數據處理、模型調用，支持可視化工作流編排	快速構建智能問答系統與複雜問答場景	適合需要快速部署知識庫問答系統的用戶	混合索引 (Elasticsearch + FAISS)、支援 PDF 表格解析 (PyMuPDF 集成)、BM25 權重可調、需人工標註種子數據
Coze	AI 智能體開發平台	整合插件、長短期記憶、工作流、卡片等功能，支持多平台發布	低門檻快速搭建個性化或具備商業價值的智能體	適合無需編程即可創建智能體的用戶	自研分佈式索引、電商數據結構

工具

- MCP
 - 2025-04-15：[automcp](#)：將任何代理、工具或編排器秒設定 MCP 伺服器
 - 2025-04-10：[line-bot-mcp-server](#)
 - 2025-04-05：[GitMCP](#)：GitMCP 太神了！一行URL 讓AI 秒懂你的GitHub 項目
 - 2025-03-14：[playwright-mcp](#)：MCP-Playwright：AI自動化神器
 - Browser-use
 - 2025-06-04：[workflow-use](#)：基於AI 的瀏覽器自動化工具：一次錄製，永久重複使用
 - 2025-04-16：[browser-use/web-ui](#)；如何使用
 - 2025-03-28：[browser-use-webui](#)
 - 2025-02-16：[browser use webui](#)部署 (實現瀏覽器自動化)
 - 2025-01-23：[Browser Use](#) – 讓AI 像人類一樣使用瀏覽器
-
- 2025-07-30：[langextract](#)；隆重推出 LangExtract：由 Gemini 驅動的資訊擷取庫
 - 2025-07-26：[presenton](#)；一款可本地部署的開源AI PPT項目，一鍵生成精美PPT
 - 2025-07-03：[MultiAgentPPT](#)；A2A+ADK+MCP多智能體並發系統產生(可線上編輯) 的PPT (含原始碼)
 - 2025-06-28：[docext](#)：基於Qwen2.5VL的文檔解析工具

- 2025-06-10 : [Agentic-Doc](#) ; [LandingAI](#)開源神器，這個Python庫讓百頁文檔秒變結構化資料！
- 2025-06-06 : [daily-arXiv-ai-enhanced](#) : 每日自動爬取arXiv論文並以LLM產生中文摘要
- 2025-05-22 : [AingDesk](#) ; [AingDesk](#) : 零門檻本地 AI 部署
- 2025-05-20 : [news-agents](#)
- 2025-05-16 : [Follow](#) ; 連續登頂GitHub 的資訊聚合神器 : [Follow](#)，讓你不再錯過任何重要資訊！
- 2025-05-11 : [SurfSense](#) : GitHub 開源專案 打通Notion、GitHub、搜尋引擎的AI超腦，如何使用
- 2025-04-28 : [PaperCoder](#) ; [Paper2Code: Automating Code Generation from Scientific Papers in Machine Learning](#) ; [PaperCoder](#) : 利用大型語言模型自動產生機器學習論文程式碼的框架
- 2025-04-16 : [OneFileLLM](#) : 這個開源神器終結了AI資料整合的惡夢！一鍵聚合網頁、程式碼、論文到剪貼簿！
- 2025-04-16 : [ScrapeGraphAI](#) : [ScrapeGraphAI](#) - 自然語言驅動的智慧爬蟲革命！
- 2025-04-15 : [stagehand](#) : [Stagehand](#) : AI驅動的下一代瀏覽器自動化框架
- 2025-04-11 : [nanobrowser](#) : AI 驅動的瀏覽器自動化神器
- 2025-04-10 : [DevDocs](#) : 開發者的文檔收割機來了！這個開源工具讓你一小時幹完一週的活！
- 2025-04-06 : [sqlchat](#) : 這款開源神器讓資料庫管理像聊天一樣簡單！
- 2025-03-26 : [pdf-craft](#) : PDF秒轉Markdown/EPUB
- 2025-03-25 : [OCRmyPDF](#) ; [OCRmyPDF](#) 能力分析
- 2025-03-12 : [AingDesk](#) ; 零門檻本地AI 部署，[DeepSeek](#)、[Llama](#) 一鍵直達！
- 2025-03-08 : [composio](#) : AI助理效率神器！[Composio](#)幫你輕鬆整合200+工具
- 2025-02-25 : [PySpur](#) : 拖曳式開發AI工作流程！這款開源神器，讓LangChain也黯然失色！
- 2025-01-13 : [DocAligner](#) : 拍照文件復原：校正扭曲、光照陰暗、版面定位
- 2025-01-13 : [PPTAgent](#) ; 中科院開源AI工具，自動將文件轉化為高品質PPT
- 2025-01-07 : [activepieces](#) : 一個開源的AI自動化工作流程工具
- 2024-12-15 : [markitdown](#)

AI Agent 開源框架

- 從AI Agent到Agent工作流程，一文詳細了解代理程式工作流程
- 萬字長文，帶你綜觀大模型Agent，涉及研究痛點、應用場景、發展方向
- 什麼是「Agentic 工作流程」？
- 什麼是Agentic AI？什麼是Agentic Workflow？與AI Agent有什麼區別和關聯？
- [FinRobot](#) ; [DeepWiki](#) ; 可支援 Gemini-2.5-Pro-preview-05-06，基於 AutoGen
- [Jupyter-AI](#) ; 可支援 Gemini-2.5-Pro-preview-05-06

-
- 2025-07-03 : [multi-modal-researcher](#)
 - 2025-06-25 : [Gemini CLI](#) : [Gemini CLI](#) : 你的開源 AI 代理
 - 2025-06-06 : [PandaWiki](#) ; 新一代AI 大模型驅動的開源知識庫建立系統
 - 2025-06-03 : [Gemini Fullstack LangGraph](#) ; DEMO ; 以為Google只是簡單放個Demo，哪想到是「開源版」Perplexity！
 - 2025-06-03 : [Perplexica](#) ; [Perplexity AI](#)，開源替代品
 - 2025-06-02 : [Paper2Poster](#) : 自動為論文產生海報
 - 2025-06-01 : [Agent Zero](#) ; 官網 ; 這個自動AI代理可以做任何事！（產生APP、程式碼、RAG 等）
 - 2025-05-30 : [WebDancer @ Alibaba](#) ; [DeepWiki](#)
 - 2025-05-28 : [Lemon AI](#) ; 全球首款全端開源通用AI Agent，讓人可以單機部署超級智慧體
 - 2025-05-25 : [OpenHands](#) ; [Demo](#)
 - 2025-05-18 : [Agent-Squad](#) ; 輕量級開源AI多智能體框架！智慧路由+上下文管理，前後端介面支援！

- 2025-05-10 : [FlowGram](#) : 字節跳動把Coze 核心開源了！視覺化工作流程引擎FlowGram 上線；如何使用
- 2025-05-10 : [DeerFlow](#) : 字節跳動DeerFlow深度解析；如何使用
- 2025-05-09 : [OpenDeepWiki](#) : 開源的DeekWiki加入MCP，輕鬆讓AI掌握開源專案使用文件！；如何使用
- 2025-05-07 : [AI Manus](#)
- 2025-04-24 : [suna](#) : 3週時間，就打造出Manus開源平替！貢獻原始碼，免費用
- 2025-04-22 : [釦子空間 \(Coze Space\)](#) : 字節版Manus 釦子空間來了！實測效果絕佳，但還有3 個問題
- 2025-04-03 : [AutoAgent](#) : 一句話全自動創建AI智能體，港大AutoAgent打造開源最強Deep Research
- 2025-04-03 : [Agent Development Kit \(ADK\)](#) : 谷歌發表「智能體開發工具包」ADK，來嚐個鮮
- 2025-04-03 : [Deepsite](#) ; DeepSite基於DeepSeek的網頁開發智能體，效果非常不錯
- 2025-03-30 : [DeepGemini](#) : AI界的'搭積木'神器，10分鐘打造你的專屬智慧團隊！
- 2025-03-24 : [AgenticSeek](#) : 又一個“Manus”開源，完全本地化替代品AgenticSeek
- 2025-03-11 : [autoMate](#) : autoMate:基於OmniParser 所建構的革命性AI自動化助手
- 2025-03-10 : [OpenManus](#) : 一文讀懂：OpenManus
- 2025-02-28 : [MoneyPrinterTurbo](#) ; AI自動生成高清短視頻
- 2024-02-01 : [MobileAgent](#) : 一句指示幫你操作手機，最新多模態手機助理Mobile-Agent來了！
- 2025-01-03 : [smolagents](#) : 新年禮物，Huggingface捲了一個Agent專案開源
- 2024-09-23 : [STORM](#) ; STORM : 一个基于LLM的知识整理系统
- 2024-10-26 : [OmniParser](#) ; 控制電腦手機的智慧體人人都能造，微軟開源OmniParser

混合專家(Mixture of Experts, MoE)模型

- 2024-12-13 : [DeepSeek-VL2](#) : DeepSeek-VL2開源，VLM邁入MoE時代，DeepSeek-VL2 先進視覺語言模型，在多模態理解方面取得了顯著進展
- [騰訊混元](#) : 騰訊混元又來開源，一出手就是最大MoE大模型
 - 2024-11-06 : [DEMO](#)
 - 2024-11-06 : [MODEL](#)

小型語言模型

- 2025-01-07 : [Smolagents](#) : Hugging Face開源全新AI智能體框架支援工具呼叫與程式碼執行！
- 2024-12-13 : [Phi-4](#) : 以小博大，微軟Phi-4正式發表~
- 2024-11-18 : [MobileLLM-1.5B](#) : Meta MobileLLM：深度架構與最佳化技術打造的行動裝置超強語言模型
- 2024-11-04 : [SmolLM2](#) : <https://github.com/hiyouga/LLaMA-Factory>
 - [Hugging Face](#)公布手機執行的小型語言模型SmolLM2
- 2024-09-25 : [Llama 3.2 90b, 11b, 3b, 1b: Revolutionizing edge AI and vision with open, customizable models](#)

推理模型

- 2025-08-05 : [gpt-oss](#) ; 隆重介紹 [gpt-oss](#) ; OpenAI重新開源！深夜連發兩個推理模型，o4-mini水平
- 2025-07-29 : [Llama Nemotron Super v1.5](#) ; 英偉達全新開源模型：三倍吞吐、單卡可跑，還拿下推理 SOTA
- 2025-07-27 : [OpenReasoning-Nemotron](#) ; 英偉達突然開源「數學核武」！1.5B 參數秒殺 o3，[OpenReasoning-Nemotron](#) 真有這麼猛？
- 2025-05-06 : [Llama-Nemotron](#) ; [DeepWiki](#) ; 英偉達發布Llama-Nemotron系列大模型，實現高效推理

- 2025-04-16 : [Video-R1: Reinforcing Video Reasoning in MLLMs](#) ; Github ; 影片推理R1時刻 · 7B模型反超GPT-4o · 港中文清華推出首款Video-R1

大型語言模型

- 2025-08-05 : [Claude Opus 4.1](#)
- 2024-11-23 : [Ai2 Tulu 3](#) : 這才是真 · 開源模型！公開「後訓練」一切 · 性能超越Llama 3.1 Instruct
 - DEMO : <https://playground.allenai.org/>
 - MODEL : <https://huggingface.co/allenai>
- 2024-11-9 : [Ai2 OpenScholar](#) : <https://openscholar.allenai.ai/>
- 2024-09-25 : [Llama 3.2 90b, 11b, 3b, 1b: Revolutionizing edge AI and vision with open, customizable models](#)

Embedding & Reranker

- 2025-07-14 : [gemini-embedding-001](#)
- 2025-06-05 : [Qwen3 Embedding](#) : 新一代文本表徵與排序模型
 - 2025-06-03 : [Qwen3-Embedding](#)
 - 2025-06-03 : [Qwen3-Reranker](#)

語音到語音大模型及工具套件

- [ten-agent](#)
 - 又一款王炸級的開源端對端語音模型
 - [TEN Framework](#) 入坑記
 - 使用 [TEN Agent](#) 搭建 [Conversation AI Agent](#)
- [pipecat](#) : OpenAI工程師親自修訂 : 用ChatGPT即時語音API建立應用
- 2025-07-21 : [Audio Flamingo 3 @ NVIDIA](#) ; NVIDIA 開源多模態音訊模型 [Audio Flamingo 3](#)
- 2025-05-08 : [Voila](#) ; 新型開源端對端AI 語音模型！ [Voila](#) : 195ms 超低延遲引領全雙工對話！
- [HuggingFace Speech-to-Speech](#)

視覺大語言模型 (Vision-Language model)

- 2025-05-20 : [Seed1.5-VL](#) ; [Seed1.5-VL](#) : 具有視覺增強多模態能力的高階語言模型 ; 字節跳動發布 [Seed1.5-VL](#) 視覺-語言多模態大模型 · 實測效果非常不錯
- 2025-05-12 : [nanoVLM](#)

多模態大語言模型 (Multimodal)

- [InternVL](#)
 - (CVPR 2024 Oral) [InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models](#)
 - [DeepWiki](#) ; [InternVL3](#) : 刷新開源多模態大模型效能新紀錄
- 2025-05-24 : [Dolphin](#) : [DeepWiki](#) ; 開源多模態複雜文件解析模型！ [Dolphin](#)
- 2025-05-21 : [Gemma 3n](#) ; [Preview](#)
- 2025-03-18 : [Mistral Small 3.1](#) : [Mistral](#)開源多模態小模型3.1 : 128K上下文+超低延遲 · 效能碾壓GPT-4o Mini
- 2025-03-14 : [Vision-R1](#) : [Vision-R1](#) : 激發多模態大模型的推理能力
- 2025-02-28 : [HumanOmni](#)
 - [阿里通義](#)開源業界首個第一視角大模型 · 超強的視訊理解效能！

- 首個專注於人類中心場景的多模態大模型，視覺與聽覺融合的突破！
- Phi
 - Phi-4-multimodal-instruct
 - 2025-02-27
 - 微軟首個多模態Phi-4問世，56億參數秒殺GPT-4o！LoRA華人大佬帶隊
 - 微軟發表Phi-4-Mini系列模型：小身材，大智慧 - 語言與多模態AI的新突破
 - 2024-09-12：微軟AI發布Phi 3.5 mini、MoE 和Vision
- MiniCPM
 - 2025-01-16：MiniCPM-o 2.6：流式全模態，端到端，多模態端側大模型來了！；文件
 - 2024-09-11：升級Ollama！MiniCPM-V2_6影像辨識模型上線
 - 2024-09-06：MiniCPM 3.0 開源！4B參數超GPT3.5性能，無限長文本，超強RAG三件套！模型推理、微調實戰來啦！：<https://github.com/OpenBMB/MiniCPM>
 - 2024-09-05：零碼基礎都敢去魔改MiniCPM-V了？是我飄了，也是Cursor 太強了：<https://github.com/OpenBMB/MiniCPM-V/>