

部署前

部署架構	一體機（包含GPU）
是否可以遠端連線部署	否
憑證類型	CSR 申請憑證
部署機器是否有網路傳輸權限（下載套件用）	是
部署機器 OS 版本（必須是 LTS）	Ubuntu 24.04 LTS(預設)
部署機器資料上傳方式	遠端連線 SCP 上傳
用戶登入方式	SAML
是否啟用 LDAP 做身份同步	否（預設）
是否設定系統郵箱（SMTP）	否（預設）

算力配置

GPU 型號	VRAM 大小	數量
NVIDIA L40S	48 GB GDDR6	4

設備IP

服務	注意事項	目標機器 IP	Port
MaiAgent API, 內部管理後台		172.20.201.71	
MaiAgent Admin UI			
MaiAgent Web Chat UI			
MinIO			
PostgreSQL		-	
Pgbouncer		-	
Elasticsearch		-	
Redis		-	
vLLM-GPU		172.20.201.71	

模型配置

模型名稱	是否啟用	Port
GPT OSS 120B	否	
MaiAgent-embeddings	是	8091

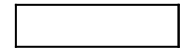
MaiAgent-rerank	是	8092
STT Parser - Whisper	是	9020
OCR Parser	是	8006

部署後

	網址	帳號	密碼
MaiAgent 內部管理後台（系統維運使用）	https://172.20.201.71:8000/admin	admin@maiagent.ai	maiagent!
MaiAgent Admin	https://172.20.201.71:3333	admin@hfcfinance.com	P@ssw0rd!
Grafana	https://172.20.201.71:3000	admin	52UfrvHSiq33IMRG4w8M

附註	
	請您們預先提供的資訊
	請您們預先開啟的對外 Port
行前確認項目	能否開放連接 USB 讓我們當下上傳檔案，或是有跳板機能讓我們 SCP。
	確認目標機器初始化（安裝作業系統）時，至少有 1TB 。
	確認 GPU 是可以調用的狀態（NVIDIA packages toolkits ⇒ nvidia-smi 能夠調用 GPU 資源）
	如果機器沒有聯網的權限，需要提前至少兩週通知準備 AirGap 套件
	配置大參數量模型再多張 GPU 上需要具備 NVLink 才可實現
	系統環境要求：Python ≥ 3.12、ansible ≥ 2.13、NVIDIA packages toolkits、Docker ≥ 19.03

預設 Port(勿修改)	domin name（需要第三方 CA 派發）
8000	
3333	
3334	
9000	
5432	-
6432	-
9200	-
6379	-
8001~8005	



	模型名稱	提供者	API Base	Value	API Key
OpenRouter	Gemini 2.5 Pro	OpenAI Compatible			
	Gemini 2.5 Flash	OpenAI Compatible			
OpenAI	GPT-5 Nano	OpenAI Compatible			
	o4-mini	OpenAI Compatible			
OpenRouter	Claude 4.5 Haiku	OpenAI Compatible			

項目	API Key
Firecrawl	

備註	
----	--

問題	回覆