按鈕：可用於新增聊天室

功能：
1.聊天介面
2.儀錶板
3.模型管理
4.設定

LyraLLM

L LyraLLM - Enterprise AI Gateway

+ 新對話

聊天

儀表板

模型管理

設定

新的對話　少於1分鐘前

執行自動路由時可查看路由決策過程

尚未開始對話

試著輸入訊息，LyraLLM 會協助你處理工作。

等待新的路由結果...

選單：可切換多個模型或自動路由

開關：開啟連網、RAG或等自定義功能

auto

Q Web Search　目 RAG Search

輸入訊息，按 Enter 送出（Shift + Enter 換行）

資料型態:String
可輸入問題並透過右側按鈕執行

可查看當前使用狀態，如**Token消耗、花費費用、請求數、回應時間等**

LyraLLM

L LyraLLM - Enterprise AI Gateway

🗨 聊天

🔲 儀表板

⚙ 模型管理

⚙ 設定

總 Token 消耗 〰	總成本 $	總請求數 🗨	平均回應時間 🕐
294,827	**$26.65**	**1017**	**317ms**
過去 7 天	過去 7 天	過去 7 天	即時監控

成本配額管理
當前配額使用情況 ($26.65 / $100)

26.6% 已使用 $73.35 剩餘

可查看本月剩餘配額

📈 **使用趨勢**
Token 消耗與成本變化

60000 ⎯ 8

45000 ⎯ 6

30000 ⎯ 4

15000 ⎯ 2

0
11月15日 11月16日 11月17日 11月18日 11月19日 11月20日 11月21日

⎯○⎯ Tokens   ⎯○⎯ 成本 ($)

模型使用分佈
各模型使用比例

GPT-4o 45%

Claude-3 12%

O1 15%

O3-mini 28%
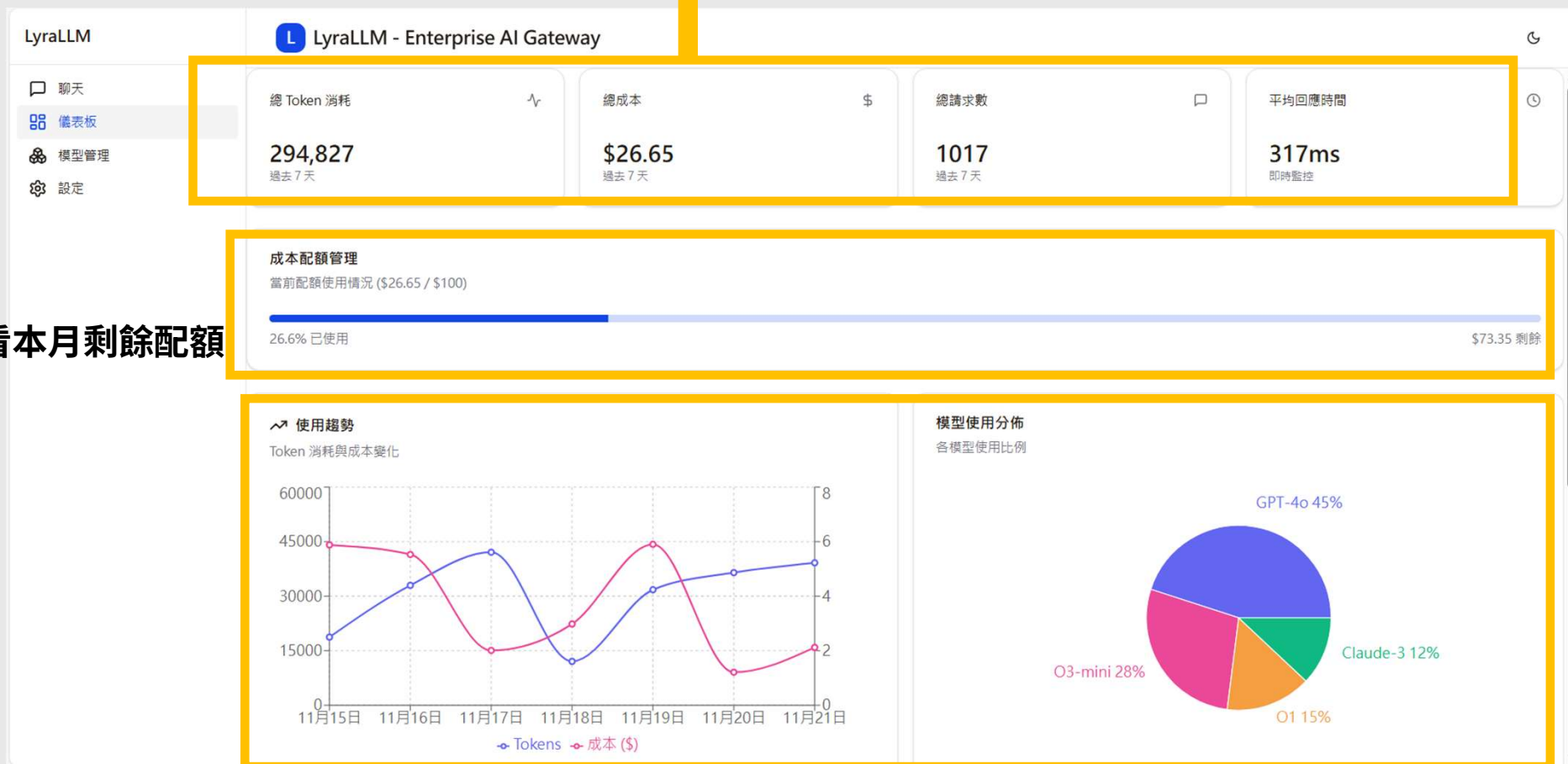
多種圖表展示用量狀況，可查看該使用者的使用趨勢、
Token變化以及模型使用分布。
後續將支援當月或當天用量圖表

LyraLLM

- 聊天
- 儀表板
- **模型管理**
- 設定

L LyraLLM - Enterprise AI Gateway

模型管理

查看和管理所有可用的 AI 模型

auto ✓
auto

Automatic routing: backend chooses the best model based on intent (select 'auto' to let router decide)

Max Tokens: 4096   Temp: 0.7

gpt-4o ✓
azure_openai

Azure OpenAI GPT-4o via Semantic Kernel

Max Tokens: 4096   Temp: 0.7

o3-mini ✓
azure_openai

Azure OpenAI o3-mini via Semantic Kernel

Max Tokens: 4096   Temp: 0.7

o1 ✓
azure_openai

Azure OpenAI o1

Max Tokens: 4096   Temp: 0.7

gpt-oss:20b ✓
ollama

Llama 2 via Ollama and Semantic Kernel

Max Tokens: 2048   Temp: 0.8

使用者能查看被授權可使用的模型列表
不同使用者會因各自權限不足所看到的模型也不同

可查看使用者個人資料

對管理者來說，此處可建立Config規則，設立不同職級
所能執行之功能、模型、服務或路由策略等。