智能診所技術架構圖

AI-powered Medical Clinic Platform Architecture

前端互動層 (User Interaction)

網頁/應用程式介面

病患透過前端介面輸入問題或需求，如療程查詢、病症問答或後續追蹤建議

API Gateway

前端介面呼叫後端 API，將用戶輸入傳遞至核心處理層

核心處理層 (LLM & RAG Workflow)

Gemini 系列模型

Google Vertex AI Model Garden 的 Gemini 模型，支援自然語言生成、摘要、分類與多輪對話

RAG 模組

向量資料庫與檢索增強生成，透過語意近似度查詢相關知識

問題分類代理

判斷問題類型（問答、推薦或排班），智能路由處理

📊 資料調用代理

從向量資料庫或排班系統提取相關資料

💬 回應生成代理

整合查詢結果與 LLM 回應，生成最終答案

🔍 語意查詢引擎

透過 Embedding API 進行向量化，實現語意搜尋

智能診所推薦平台

🎯 智慧推薦引擎

根據病患問題推薦個人化療程或知識文章

👤 個人化療程推薦

基於病患歷史和偏好，提供客製化治療建議

📋 診所配對系統

智能匹配最適合的診所和醫療資源

資料儲存層 (Data Layer)

🧠 向量資料庫

pgvector 或 Pinecone，儲存醫療知識的語意向量，支援快速語意查詢

🗄️ 結構化資料庫

Cloud SQL，儲存排班資料、病患記錄等結構化資訊

📚 醫療知識庫

療程資料、醫學文章等專業知識內容庫

模型評估與監控層

📈 Vertex AI Evaluation

評估 LLM 回應的準確性、一致性、完整性與毒性過濾

🔍 系統監控

Cloud Trace 與 Cloud Logging，追蹤 API 延遲、錯誤率與服務健康狀態

📊 Looker Studio

可視化 RAG 命中率與回應品質，提供即時監控儀表板

部署與服務層 (Serving & DevOps)

Vertex AI Agent Engine

Fully-managed 部署方案，適合快速上線

FastAPI / Cloud Run

自建服務，支援客製化與彈性擴充

Terraform

基礎設施即代碼，確保環境一致性

CI/CD Pipeline

Cloud Build 實現持續整合與部署流程

資料與控制流

1 病患輸入問題（前端介面） →

2 問題經 API Gateway 傳至核心處理層 →

3 問題分類代理判斷處理路徑 →

4 RAG 流程：問題向量化 → 向量資料庫查詢 → 知識片段檢索 →

5 代理協作整合 LLM 回應與檢索資料 →

6 回應返回前端，同時記錄於監控系統 ✓

前端互動層 核心處理層 推薦平台 資料儲存層

評估監控層 部署服務層