LLM-enabled Network Service Monitoring Analyzer in Edge Intelligence Environment

Students: 黃韋傑、林哲亨、林宣辰、黃教丞
Advisor: Prof. Hojjat Baghban（霍杰特）

摘要

本研究旨在邊緣運算環境下的 Kubernetes (K8s) 叢集，設計一套智慧監控分析系統。透過整合 Prometheus 監控數據、檢索增強生成技術，以及具備為運工具功能的本地端大型語言模型，我們將複雜的維運任務轉化為直觀的對話式體驗。

本系統允許維運人員透過自然語言查詢即時系統狀態，大幅降低學習複雜查詢語法（如 PromQL）的門檻。此外，研究中模型的本地化部署，確保數據的安全性與隱私，驗證 Edge AI 在智慧維運領域的可行性。進一步發展基於自然語言的 Pod 自動水平擴展 HPA 機制。

背景與問題

隨著雲原生技術普及，維運門檻日益提高

1. 複雜度高：Kubernetes (K8s) 架構龐大，傳統監控依賴 Prometheus 與 Grafana，這些工具對非技術人員並不友善。
2. 學習曲線陡：維運人員需要精通 PromQL (查詢語法)、複雜的 Kubectl 指令才能診斷問題。導致開發者與維運數據之間存在巨大的鴻溝。
3. 數據主權與外洩風險:許多新興的 AIOps 或 LLM 監控服務都建構在公有雲上。企業須將最敏感的系統日誌 、指標上傳到第三方伺服器，引發資料外洩與隱私風險

解決方案：以 LLM 為核心的智慧中樞

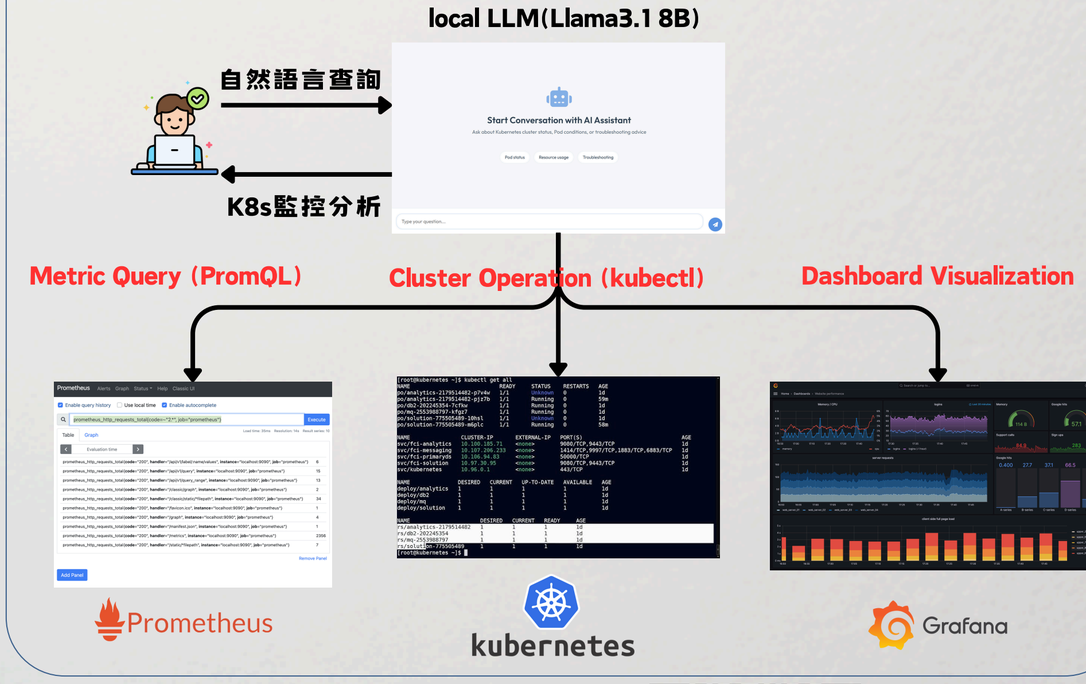
我們提出一個以大型語言模型為核心的架構，突破工具間的壁壘。

整合三大運維神器

- Prometheus: 即時撈取關鍵效能指標 (Metrics)，如 CPU、記憶體負載，精確掌握資源狀態。
- Grafana: 將抽象、複雜的數據自動轉化為直觀的圖表，讓趨勢一目了然。
- Kubectl: 執行由自然語言轉譯而來的精確指令，對 K8s 叢集的直接操作與管理。

讓使用者透過「自然語言」即可完成監控與初步維運，讓新手也能像資深運維工程師一樣管理K8s叢集。

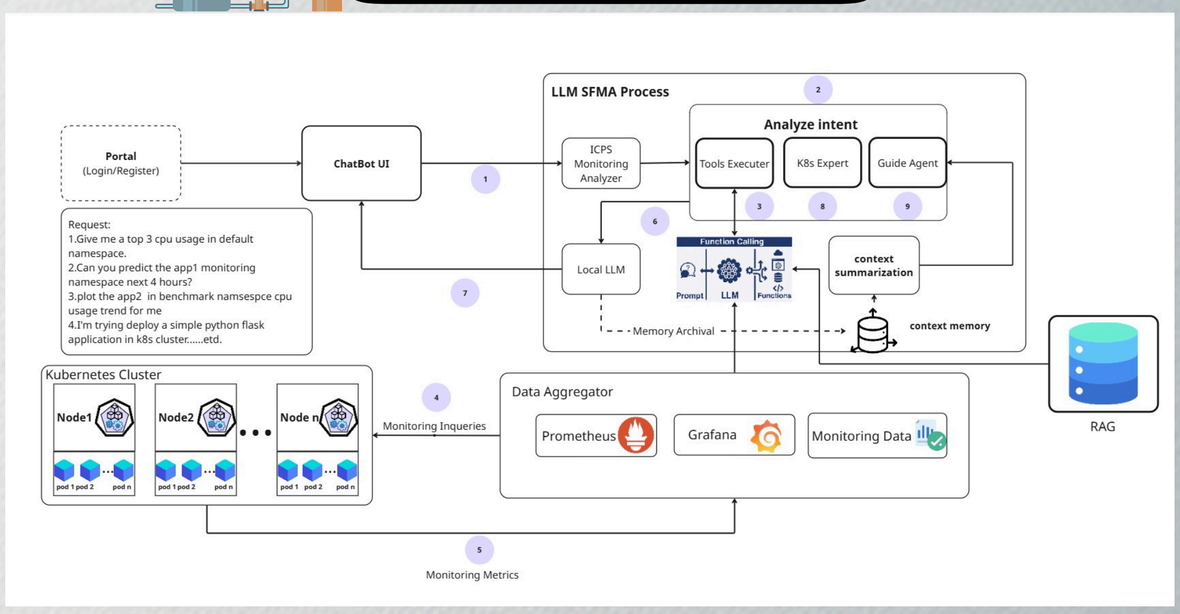
系統示意圖



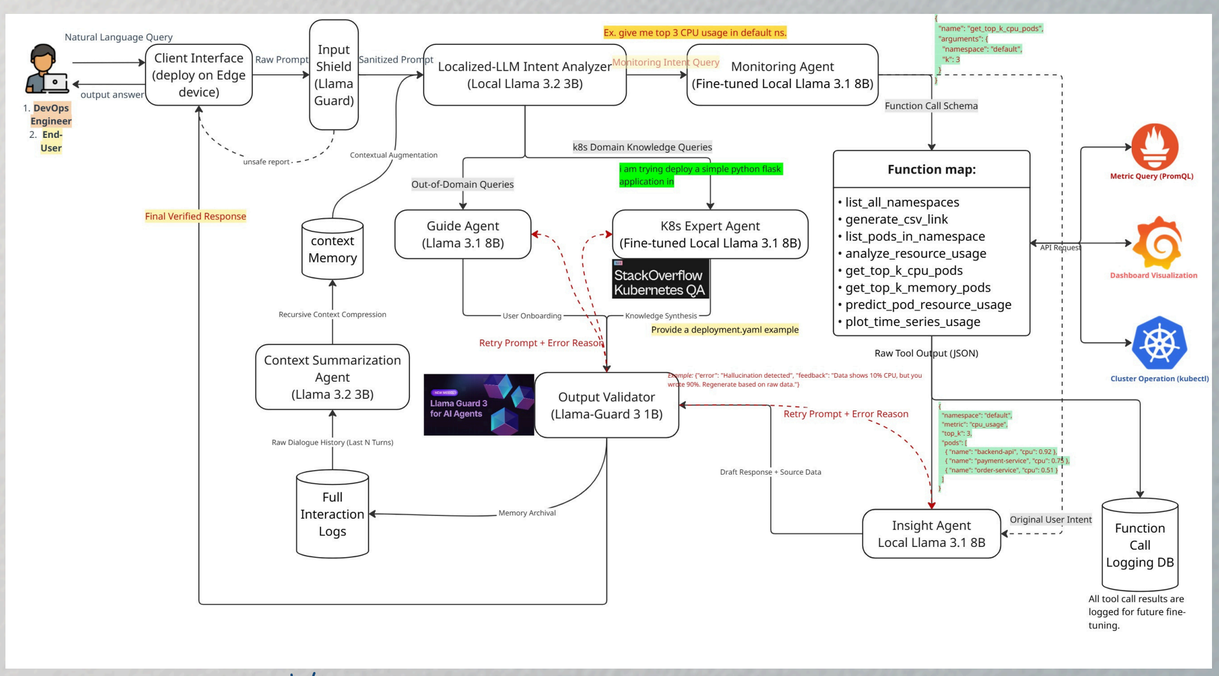
核心技術

Model: Llama 3.1 8B
Backend: FastAPI, Python
Infrastructure: Kubernetes, Prometheus, Grafana
Key Tech: Retrieval-Augmented Generation , Function Calling , ARIMA
LoRA Fine-Tuning , Edge AI Server

系統架構與成果



1. 監控： Prometheus 持續監控 K8s 叢集，收集 CPU、記憶體等數據並存入「Monitoring Data」資料庫。
2. 互動： 使用者以自然語言向微調後的本地 Analyzer 提問。
3. 呼叫: Analyzer 使用 Function Calling 解析使用者意圖，決定要呼叫哪一個工具。
4. 存取: 被 Analyzer 呼叫的工具向「Monitoring Data」資料庫請求所需數據。
5. 整合: 資料庫回傳原始數據，工具處理並轉換為簡潔的結果（如即時 CPU 值）。
6. 匯集： Function Calling 機制將處理完畢的數據封裝成結構化的 JSON 格式，並回傳給 Analyzer。
7. 回覆： Analyzer 結合收到的 JSON 數據，生成一段簡潔易懂的自然語言回覆給使用者。
8. 諮詢： 意圖為詢問概念，微調過的「K8s 專家」，提供權威解釋與操作指引。
9. 引導 ： 若指令不明確，「指導助手」會引導提問或提供選項，協助使用者補全資訊完成有效查詢。



結論與未來展望

「LLM 網路服務監控分析師」，微調一個能在本地端、邊緣裝置運行的語言模型。我們將傳統監控流程徹底改變。使用者只需用「幫我查五分鐘 CPU 使用率」這類提問，系統即會自動轉換為精確查詢，回傳資源消耗排名列表或趨勢預測等「深度分析」，徹底保障企業日誌與效能指標絕不外洩的數據隱私需求。

展望未來，系統升級為「主動維運助手」，在偵測到潛在風險時主動發出警告後提供具體建議。我們將專注於企業級與邊緣部署落地，滿足安全合規需求，規劃分級版本，精準切入智慧製造、私有雲等對數據隱私高度要求的場域。

Live System Demo
ICPS Lab Showcase
Intelligent Cyber-physical Systems Research Group,