AI-Powered Log Analysis: A Comparative Case Study

Konstantinos Soufleros

For position Junior AI engineer

May 2025

Case Study Overview

- Goal: Analyze logs with AI for prediction & correction

- Deliverables: 2 Use Cases + Bonus Proposal

- Evaluation: Accuracy, latency, explainability, setup complexity

Use Case 1: Google Cloud AI

- Vertex AI + Pub/Sub + Cloud Run
- Model: gemini-2.0-flash-lite-001
- Trigger: Error logs via Pub/Sub
- Output: LLM-based root cause explanations

GCP Pipeline Architecture

Cloud Logging → Pub/Sub → Cloud Run → Vertex AI

Each component:

- Cloud Logging: Writes log
- Pub/Sub: Routes log
- Cloud Run: Parses & sends to LLM
- Vertex AI: Diagnoses & explains

Use Case 2: Local EFK + LLM

- EFK Stack: Elasticsearch + Filebeat + Kibana
- Python script queries logs
- Local LLM: llama3.2 via Ollama
- Output: Console-based explanations

EFK + Ollama Architecture

Filebeat → Elasticsearch → Python script → Ollama

Details:

- Filebeat forwards logs
- Python script tails ES index
- Ollama interprets and explains errors

Use Case Comparison

Feature	Use Case 1: GCP Vertex AI	Use Case 2: Local EFK + LLaMA
Setup Complexity	Medium (IAM, Pub/Sub, Cloud Run)	Medium (Docker, LLM setup)
Trigger	Event-driven	Polling-based
Cost	Free Tier Eligible	Free (resource-dependent)
LLM Response Time	Fast	Slow
AI Model	Gemini-2.0-Flash	LLaMA 3.2 (Ollama)
Extensibility	Very High	High
Automation	Excellent (Workflows)	Manual scripting
Production-Readiness	✅ Yes	⚠️ Prototype
Output	Cloud Run Logs	Terminal Output

Evaluation & Observations

Use Case 1: Google Cloud Vertex AI

- Real-time analysis pipeline with serverless execution
- Scalable, cost-efficient, and suitable for production deployment
- Easy integration with cloud-native workflows and automation tools

Use Case 2: Local EFK Stack + LLaMA

- Effective for local, offline prototyping and testing
- Offers full control, no external cost, but limited in performance and scale
- Ideal for experimentation where data privacy or cloud access is restricted

Limitations

- Performance Constraints on Local Machine
- Limited Evaluation Metrics
- Basic Log Ingestion
- Simplified LLM Prompts
- Security & Cost Awareness

Future Work

- Quantitative Evaluation
- Real-Time Alerting and Remediation
- Richer LLM Capabilities
- Scalable Cloud Alternative
- Anomaly Detection Integration

Thank You

- Questions?

Konstantinos Soufleros
soufleros.kostas@gmail.com