Project Name: LLM-Assisted Incident Summarization and Clustering from System Logs

Project Proposer: Alison Yao alisonyao821@gmail.com

Open Source: Yes

Mentors: Alison Yao alisonyao821@gmail.com

Preferred Experience

Required for all members: Python, Git, Experience with LLMs

Required for at least one member: Experience with ML/NLP, Familiarity with embeddings, Experience with cloud

Valuable: Experience with LLM APIs, Experience with building data pipelines, Familiarity with clustering

Nice to have: Frontend/Dashboard capabilities to present outcome

Project Background

Modern cloud systems generate massive volumes of logs during failures and outages. In practice, raw logs are too verbose for direct analysis, and engineers rely on summarized incident views before identifying related or recurring incidents.

This project is inspired by real-world incident triage systems used in industry, where logs are first compressed into incident-level summaries and then grouped to identify duplicates and correlation, or recurring failure patterns.

Project Description

The team will build a cloud-log-based system that:

1. Ingests large-scale system or application logs from public datasets
2. Uses Large Language Models to generate structured incident summaries from raw logs
3. Converts incident summaries into embeddings
4. Clusters incidents to identify:
    - duplicate incidents
    - recurring failure patterns

The final system should expose:

- A backend service for log ingestion, summarization, and clustering
- A simple UI or API to explore incidents and clusters

Learning Outcomes

Students will gain experience in:

- Designing scalable cloud data pipelines
- Working with real-world, noisy log data
- Applying LLMs for summarization and structured extraction
- Using embeddings and clustering for similarity analysis
- Making tradeoffs between accuracy, cost, and latency in cloud systems
- Communicating operational insights from unstructured data