# AI-Powered Data Analytics for IT Operations

**PhD Topic:**

My research focuses on leveraging **data science, machine learning, and large language model (LLM)** techniques to analyze and optimize **IT system operations**.

**Research Domains:**

1. HPC & Operational Data Analytics
2. LLM Service Availability
3. AI for IT Operations (AIOps)

**Xiaoyu Chu**

chuxiaoyu123@gmail.com

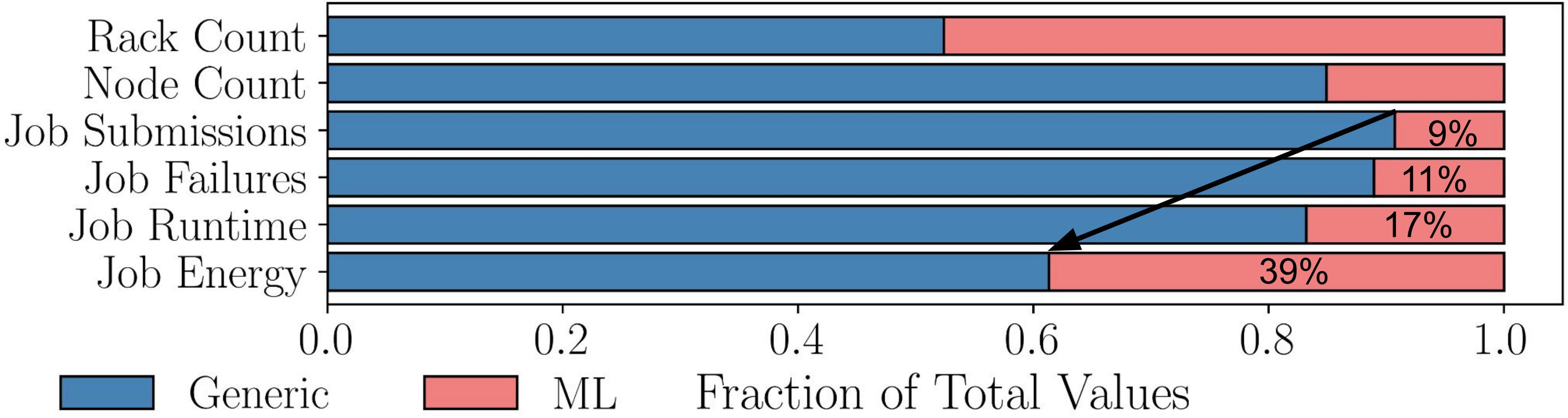
https://chuxiaoyu.github.io/

# Roadmap

| # | Domain(s) | Projects | Focus | Main Result |
|---|---|---|---|---|
| 1 | HPC & Operational Data Analytics | 1. ML Workload Characterization (HotCloudPerf’23, ICPADS’24)<br>2. ODAbler (GraphSys’24) | 1. Characterized over 1.6 million jobs using performance, failure, and power data from a production HPC datacenter (SURF Lisa).<br>2. Formalize heterogeneous operational metrics into a ontology-based graph structure (82 classes, 63 data properties). | 1. Unsuccessful job terminations consumed approximately 50% of the total cluster energy,. ML jobs consumed 39% of energy despite only 9% of total submissions.<br>2. The resulting graph structure enables complex relational queries (e.g., SPARQL) that link isolated metrics across the hardware-software stack. |
| 2 | LLM Service Availability | 3. Empirical Analysis of LLM Outages (ICPE’25)<br>4. FAILS Framework (HotCloudPerf’25) | 1. Conducted empirical characterisation of over 500 incidents across 8 LLM services using the industry failure recovery model (MTTR/MTBF).<br>2. Designed and implemented FAILS, the first open-sourced framework for automated LLM incident collection, analysis, and visualization. | 1. Anthropic services showed a severe lack of failure isolation, with 71.79% of incidents impacting all monitored services simultaneously. LLM service failures exhibit strong weekly and monthly periodicity.<br>2. FAILS successfully integrates LLM-assisted components (chatbot/plot analysis) to provide contextual insights into complex failure patterns and statistics. |
| 3 | AI for IT Operations (AIOps) | 5. LLM-powered Incident Report Data Extraction (Under Review) | 1. Developed a methodology to automate extracting structured information (10 fields) from >3,000 unstructured Cloud Incident Reports (AWS, Azure, GCP) using 6 LLMs and prompt engineering. | 1. Lightweight LLMs (Gemini 2.0 / GPT 3.5) offered the best performance-cost trade-off, achieving competitive accuracy while being 50–60× less expensive than SotA models. |

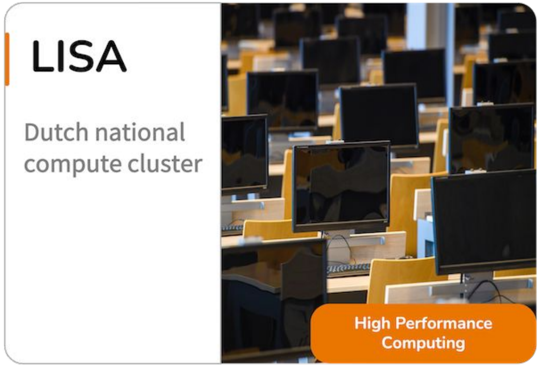
# D1 - Project: ML Workload Characterization

**Research Question:** How do Machine Learning (ML) workloads impact the performance, failures, resource utilization, and energy consumption of HPC datacenters compared to generic, compute-intensive workloads?



# D1 - Project: ML Workload Characterization

**Research Approach:** This work leveraged long-term operational data collected from a national-scale production HPC datacenter (SURF Lisa). The methodology involved integrating approximately 1.6 million job records (from SLURM), and 128 million node data records (from Prometheus) into a joint dataset. Statistical techniques were used to compare workload characteristics, energy usage, and job exit states.



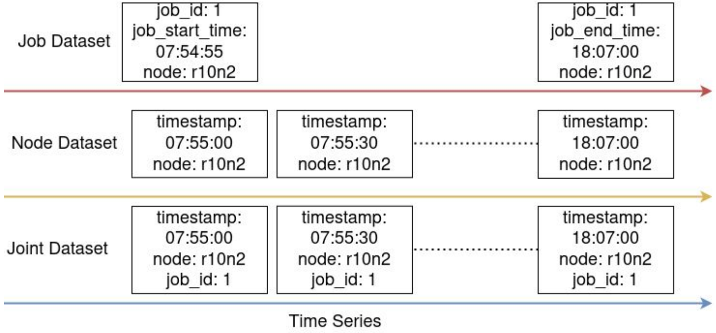
Scheduler:

slurm
workload manager

Monitoring:

Prometheus

Dataset Integration:



Job Dataset

job_id: 1
job_start_time:
07:54:55
node: r10n2

job_id: 1
job_end_time:
18:07:00
node: r10n2

Node Dataset

timestamp:
07:55:00
node: r10n2

timestamp:
07:55:30
node: r10n2

timestamp:
18:07:00
node: r10n2

Joint Dataset

timestamp:
07:55:00
node: r10n2
job_id: 1

timestamp:
07:55:30
node: r10n2
job_id: 1

timestamp:
18:07:00
node: r10n2
job_id: 1

Time Series

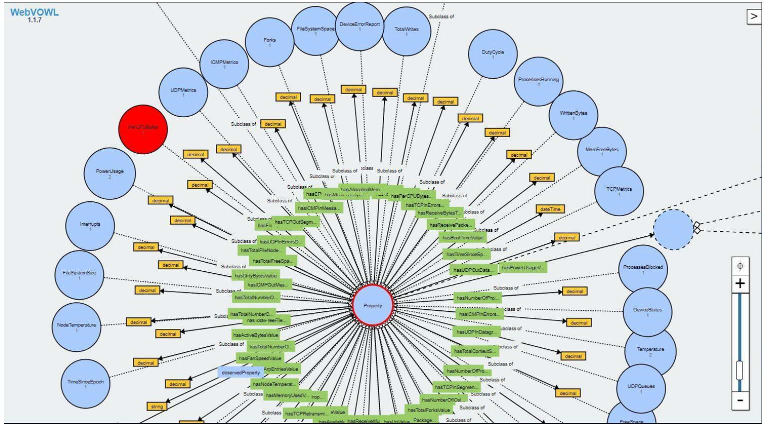
# D1 - Project: ML Workload Characterization

**Main Scientific Result:** Approximately 50% of the total cluster energy was consumed by jobs that terminated unsuccessfully (e.g., failures, timeouts, out-of-memory errors). Furthermore, ML jobs demanded a disproportionately large 39% of the cluster-wide energy budget, despite representing only 9% of total submissions.

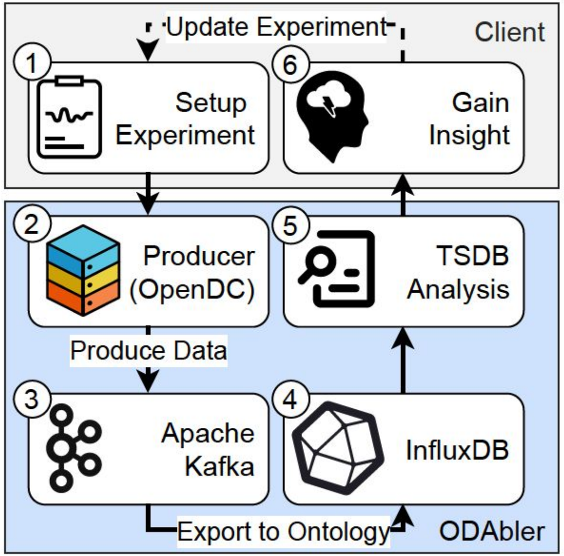
**Actionable Insight:** Checkpointing mechanisms should be implemented to save partial work, which helps avoid the substantial energy wastage caused by unsuccessful job terminations. Also, prioritizing GPU resource allocation at positions with superior cooling can enhance performance, as GPU temperatures regularly reached critical limits.

# D1 - Project: ODAbler - Ontology Design for Operational Data Analytics

**Research Question:** How can the massive and heterogeneous operational data from datacenters be organized, shared, and leveraged using a graph-based ontology to enable more effective Operational Data Analytics (ODA)?



Graph structure.



High-level architecture of ODAbler framework.

# D1 - Project: ODAbler - Ontology Design for Operational Data Analytics

**Research Approach:** This project focused on designing and implementing a graph-based ontology. Operational metrics collected from HPC clusters were mapped into a formal graph structure using OWL. This design process culminated in the design of the ODAbler framework to ingest and export operational metrics for further analysis and simulation.

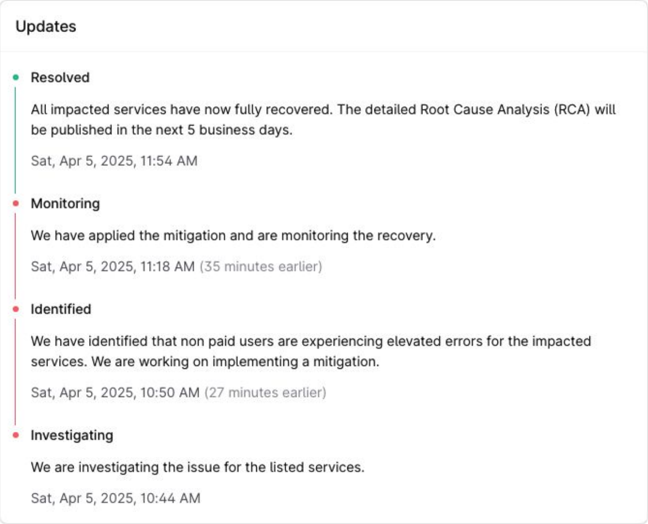
**Main Scientific Result:** The designed ontology structure successfully formalized complex hierarchies (82 classes, 17 object properties, 63 data properties) that capture datacenter operational status, establishing a viable foundation for organizing ODA data in a graph format.

**Actionable Insight:** Using a common, ontology-normalized data structure facilitates the use of powerful graph-aware query engines (e.g., SPARQL), allowing operators to link isolated metrics across the hardware-software stack for deep analysis.

# D2 - Project: Empirical Analysis of LLM Outages

**Research Question:** What are the empirical characteristics of outages and failure-recovery processes in public Large Language Model (LLM) services?

In a significant disruption, OpenAI's popular AI chatbot, ChatGPT, experienced a global outage on Thursday morning, leaving millions of users unable to access its services for nearly three hours.



Updates

- Resolved

  All impacted services have now fully recovered. The detailed Root Cause Analysis (RCA) will be published in the next 5 business days.

  Sat, Apr 5, 2025, 11:54 AM

- Monitoring

  We have applied the mitigation and are monitoring the recovery.

  Sat, Apr 5, 2025, 11:18 AM (35 minutes earlier)

- Identified

  We have identified that non paid users are experiencing elevated errors for the impacted services. We are working on implementing a mitigation.

  Sat, Apr 5, 2025, 10:50 AM (27 minutes earlier)

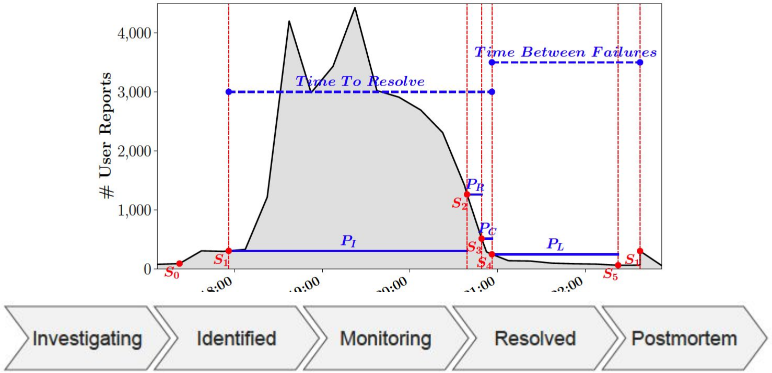
- Investigating

  We are investigating the issue for the listed services.

  Sat, Apr 5, 2025, 10:44 AM

# D2 - Project: Empirical Analysis of LLM Outages

**Research Approach:** This study collected long-term datasets of outages and incidents from 8 LLM services across 3 major providers (OpenAI, Anthropic, and Character.AI). The analysis used an industry-standard failure recovery model to calculate dependability metrics such as Mean Time To Recovery (MTTR) and Mean Time Between Failures (MTBF). The work also included analysis of temporal patterns and failure co-occurrence.



Investigating › Identified › Monitoring › Resolved › Postmortem

Key parameters and metrics:

- $P_I$: Investigating Period
- $P_R$: Repairing Period
- $P_C$ : Checking Period
- $P_L$: Learning Period
- MTTR: Mean Time To Resolve
- MTBF: Mean Time Between Failures

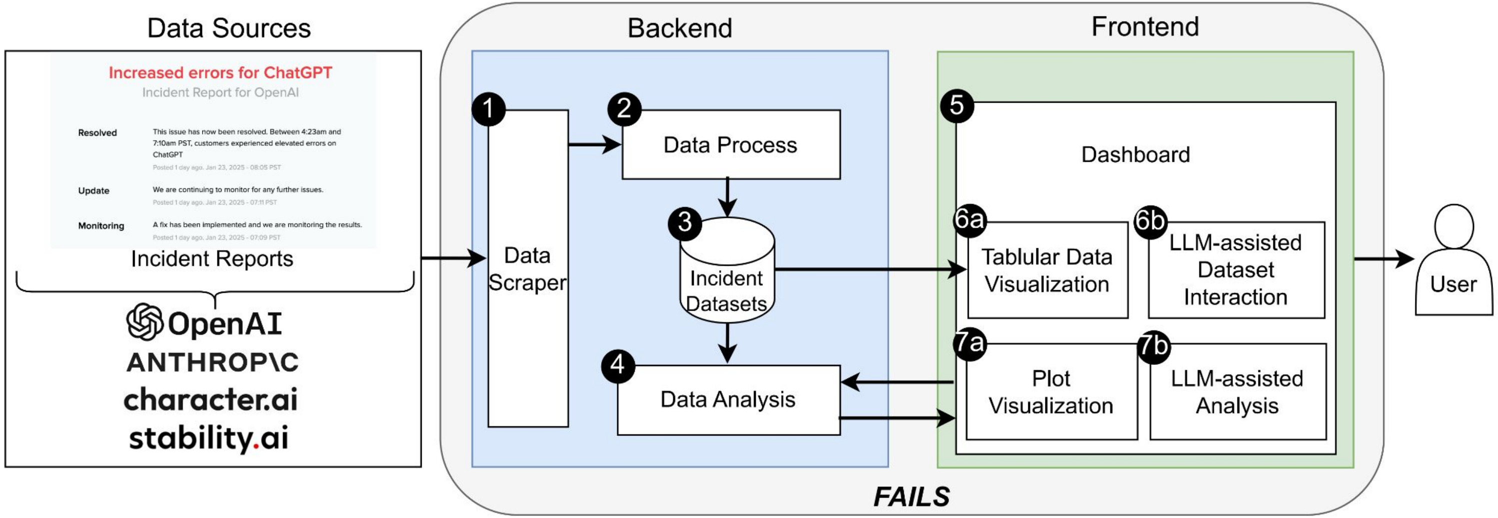
# D2 - Project: Empirical Analysis of LLM Outages

**Main Scientific Result:** Anthropic services exhibited significantly less failure-isolation, with the majority of incidents (71.79%) affecting all its monitored services simultaneously. In contrast, OpenAI incidents tended to impact only a single service (57.63%). Strong weekly and monthly periodic patterns were observed in LLM service failures.

**Actionable Insight:** LLM providers must improve service isolation mechanisms to prevent internal failures from cascading and affecting multiple services simultaneously. Additionally, operators should employ faster testing and continuous deployment techniques to reduce MTTR, as the majority of resolution time is spent in the Checking period.

# D2 - Project: FAILS Framework

**Research Question:** How to design a tool which can automatically collect, analyze, and interpret incident reports and failure patterns for different LLM services and providers?



**Increased errors for ChatGPT**

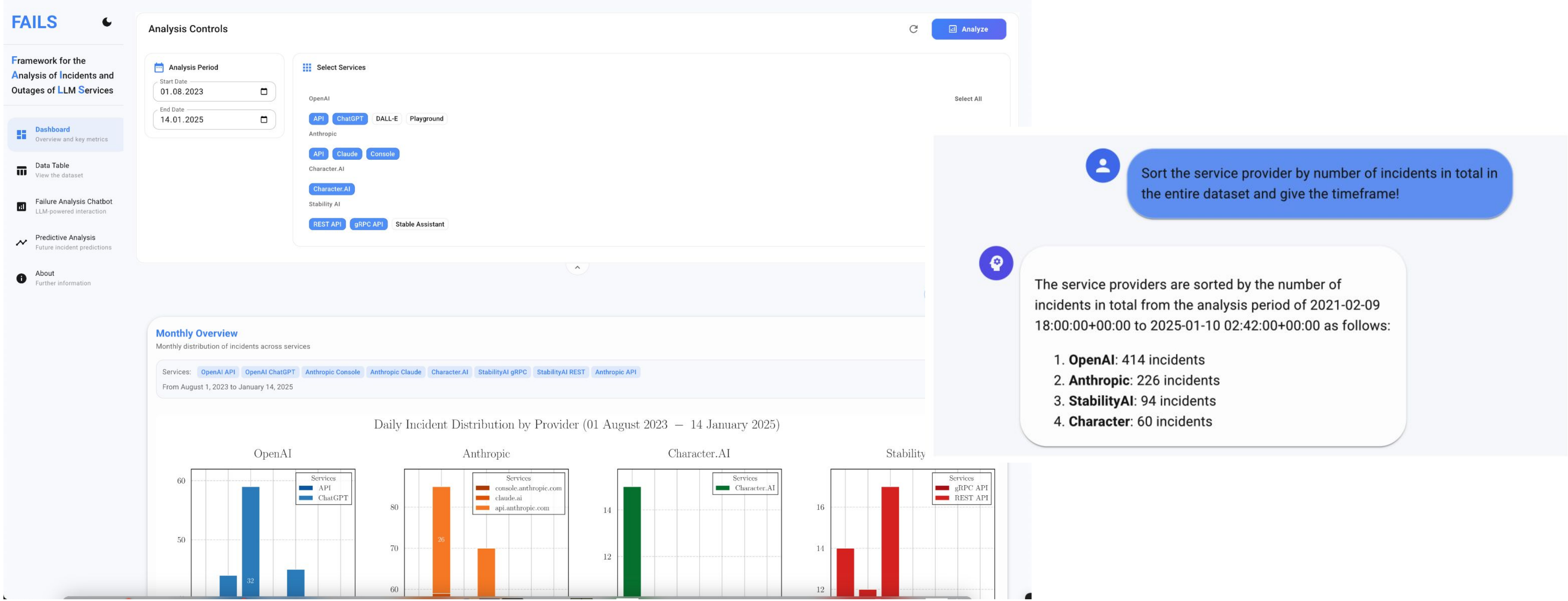
Incident Report for OpenAI

| | |
|---|---|
| **Resolved** | This issue has now been resolved. Between 4:23am and 7:10am PST, customers experienced elevated errors on ChatGPT<br>Posted 1 day ago. Jan 23, 2025 - 08:05 PST |
| **Update** | We are continuing to monitor for any further issues.<br>Posted 1 day ago. Jan 23, 2025 - 07:11 PST |
| **Monitoring** | A fix has been implemented and we are monitoring the results.<br>Posted 1 day ago. Jan 23, 2025 - 07:09 PST |

Architecture Overview.

# D2 - Project: FAILS Framework

## Front-end Interactions of FAILS:



FAILS

Framework for the Analysis of Incidents and Outages of LLM Services

Start Date
01.08.2023

End Date
14.01.2025

OpenAI
API ChatGPT DALL-E Playground

Anthropic
API Claude Console

Character.AI
Character.AI

Stability AI
REST API gRPC API Stable Assistant

**Monthly Overview**

Monthly distribution of incidents across services

Services: OpenAI API OpenAI ChatGPT Anthropic Console Anthropic Claude Character.AI StabilityAI gRPC StabilityAI REST Anthropic API

From August 1, 2023 to January 14, 2025

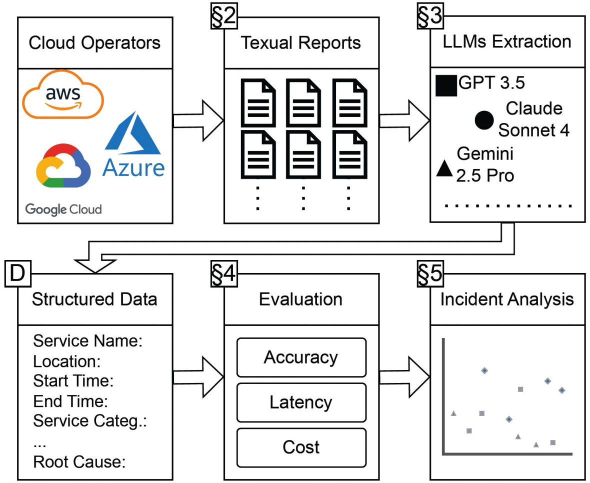
Sort the service provider by number of incidents in total in the entire dataset and give the timeframe!

The service providers are sorted by the number of incidents in total from the analysis period of 2021-02-09 18:00:00+00:00 to 2025-01-10 02:42:00+00:00 as follows:

1. **OpenAI**: 414 incidents
2. **Anthropic**: 226 incidents
3. **StabilityAI**: 94 incidents
4. **Character**: 60 incidents

# D3 - Project: LLM-based Extraction from Cloud Incident Reports

**Research Question:** How to apply accurate and cost-efficient are state-of-the-art LLMs and prompt engineering strategies for extracting structured information from complex, unstructured cloud incident reports to enable long-term dependability analysis?

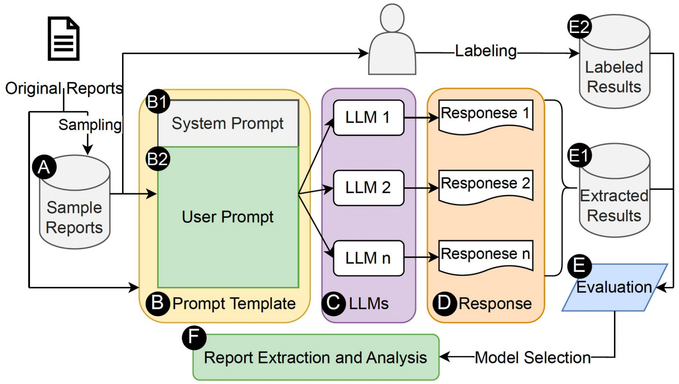


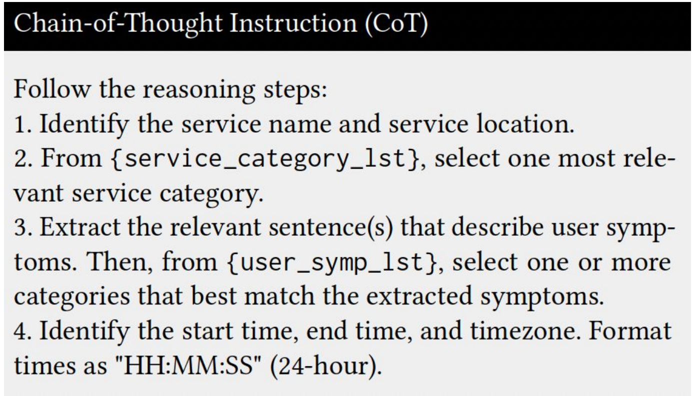
Prompt Strategies.

| Label | Components |
|---|---|
| Full-ZS | Task + CoT + Category + Format |
| Full-FS | Task + CoT + Category + Examples + Format |
| Basic-ZS | Task + Format |
| Basic-FS | Task + Examples + Format |
| CoT-ZS | Task + CoT + Format |
| Categ-ZS | Task + Category + Format |

# D3 - Project: LLM-based Extraction from Cloud Incident Reports

**Research Approach:** This work collected over 3,000 cloud incident reports (AWS, Azure, GCP) and manually annotated 460 reports to establish ground truth. A methodology was designed to systematically compare six prompt strategies (e.g., Few-shot/Zero-shot, CoT) and six LLM models (lightweight and SotA) to extract ten types of structured information (e.g., root cause category, service name). Evaluation measured accuracy (Exact Match, Token-level F1, BERTScore), latency, and token cost.





Chain-of-Thought Instruction (CoT)

Follow the reasoning steps:
1. Identify the service name and service location.
2. From {service_category_lst}, select one most relevant service category.
3. Extract the relevant sentence(s) that describe user symptoms. Then, from {user_symp_lst}, select one or more categories that best match the extracted symptoms.
4. Identify the start time, end time, and timezone. Format times as "HH:MM:SS" (24-hour).

# D3 - Project: LLM-based Extraction from Cloud Incident Reports

**Main Scientific Result:** LLMs achieved high metadata extraction accuracy (75%–95%), and Few-shot prompting generally improved accuracy for metadata fields, sometimes by an average of 17.34%. Statistical analysis of the LLM-extracted data showed that the most frequent root causes were DEPLOY (35.43%), CONFIG (18.11%), and EXTERNAL (14.96%).

**Actionable Insight:** Lightweight models such as Gemini 2.0 and GPT 3.5 offer a strong balance of accuracy, cost, and latency, achieving competitive performance while being 50–60× less expensive than top-tier models for extraction tasks. It is recommended to employ Few-shot prompts and begin with lightweight models when integrating LLMs into AIOps extraction pipelines.

# Further Information

**Links**

- Homepage: https://chuxiaoyu.github.io/
- GitHub: https://github.com/chuxiaoyu
- Google Scholar: https://scholar.google.com/citations?user=eIKz6sMAAAAJ&hl=en
- Linkedin: https://www.linkedin.com/in/chuxiaoyu/

# Work In Process (WIP) and Future Work

## Work In Process (WIP)

- LLM Uptime Archive
-

## Future Work

- AIOps
- MLOps
- LLMOps