Computer Fraud and Security

ISSN (online): 1873-7056

Incident Intelligence in Telecom: A Framework for Real-Time Production Defect Triage and P0 Resolution

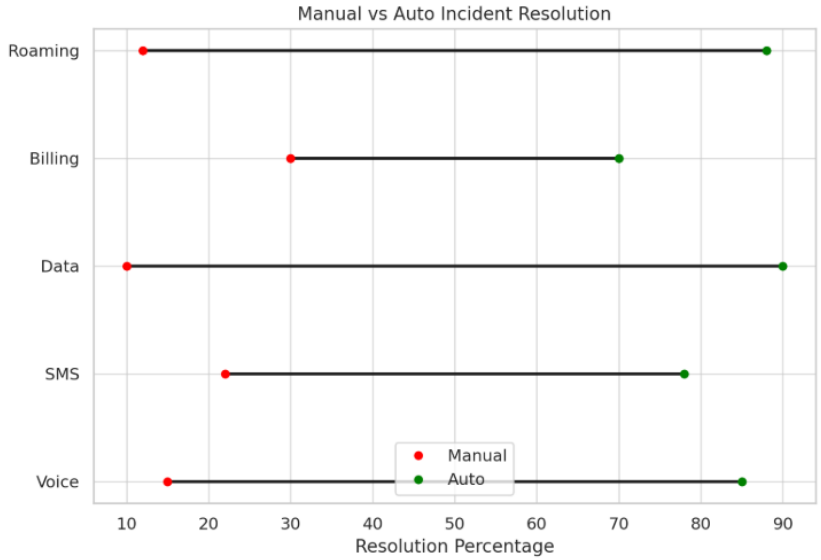
Suresh Kumar Panchakarla

ABSTRACT: The rising intricacy of the telecom platforms drives the need of the intelligent, automated incident response systems. The current paper describes a real-time incident intelligence platform that was implemented into Charter Communications Mobile 2 ecosystem. Using Kafka-based ingestion of logs, the machine learning logic of chooser responder, and RCA pipelines that are automatic with Splunk and Datadog, this framework will lower the mean time to detect, assign, and resolve P0 incidents considerably. The deployment in the real world illustrates the better levels of keeping to the SLA, automation of the triage and resilience of the systems. The architecture will combine well-organized playbooks and feedback loops to permit continuous learning. The findings indicate that these structures can be the framework to provide a model of scalable, intelligent triage of production defects in a telco-grade application.

KEYWORDS: Telecom, P0, Production, Incident Intelligence

I. INTRODUCTION

Major production flaws in the telecommunication infrastructures are becoming severe and require smart decision towards immediate solution. The classical manual triaging procedures are associated with delays, regression of service, and dissatisfaction of customers. In this paper, a framework in the form of incident intelligence is proposed, which would help to speed up the recognition process, automate triage and streamline resolution processes in telecom environments.



The solution is constructed on top of the Mobile 2 platform of Charter Communications to combine streaming data (transiting through Kafka) and observability tools (Splunk, Datadog) and uses machine learning to perform root cause analysis and predictive engagement by responders. Automated and based on structured operation playbooks in telco-grade networks, the framework is a response to the most topical concern of the need to manage incidents faster and scalable.

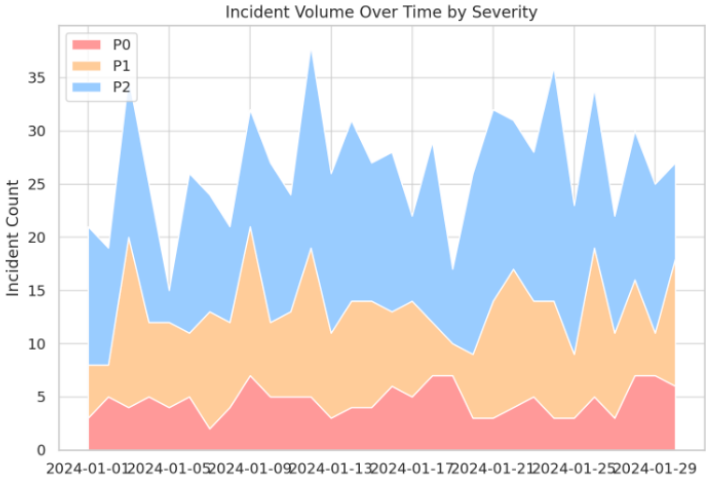
II. RELATED WORKS

Automated Triage

The field of incident triage has taken a new turn in telecommunication networks where automation and machine learning have become the rapidly relied themes towards swift solutions to serious problems in the production

lines. Manual approaches to the subject of triage have been rapidly transitioning to automated models that are capable of processing past information and assigning the most appropriate personnel within seconds.

New advances in incident triage systems are concerned with how to take advantage of ordered incident histories, combining human patterns of resolution, and exploiting machine learning to prioritize similar previous incidents that allow superior allocator determination and even resolution mapping [1].



The method not only brings about high triage times but also minimizes the use of institutional knowledge and tribal memory, which is an essential drawback in the case of high churn in the telecom operational teams. Alarm root cause analysis is very useful in the aspect of network preservation.

The very nature and number of produced alarms in the contemporary telco networks require sophisticated analytical methods. There exists further sophistication of root cause identification under hybrid modeling in causal inference model with network embedding methods, such as Hawkes Process-based Causal Inference (HPCI) and Causal Propagation-Based Embedding (CPBE) [2].

Compared with more traditional co-occurrence or correlation-based methods applied on real-time data pipelines, these models are more accurate and assist in the creation of incident graphs that would continuously change depending on the dynamic system behavior. The new influence maximization techniques will take care of not only making sure that all the alarms can be taken care of within a short time but also explain them in the context of the larger interdependent network topology which is also a vital necessity in the telco scale ecosystems.

In microservices-based systems more typical of modern network infrastructure like a telecommunications company, where the scale is so large that the incident management of the problem is out of its scale, solutions to incident at scale, e.g. through event graph-based root cause analysis frameworks like Groot proved incident management at scale to be possible.

The systems use heterogeneous data streams logs, traces and metrics and turn them into causality graphs which identify sources of failure with impressive accuracy, up to 95% top-3 RCA accuracy in production-grade systems [3]. Relevance and interpretability of results are also provided by the incorporation of domain-specific rules and customization by SREs (Site Reliability Engineers), which may also become a necessity in the case of P0 outages when the human authorization may be a crucial method even in the case of automation.

AIOps Foundations

Artificial Intelligence of IT operations (AIOps) has become one of the fundamental drivers of smart incident identification and recovery, specifically in the high-changeover telecom deployments. Static rules and other threshold-based notification systems cannot be used to cover the dynamically changing workloads and streams of data in telco-grade systems.

AIOps uses both big data analytics and machine learning to identify anomalies, correlate alerting, and self-heal in time to triage, mitigate in real-time [4]. The disarray that plagues the field, in addition to the absence of universally agreed standards or a deployment model, creates disparities in implementation in each organization.

Attempts to bring standard definitions of taxonomies and structured basis models of operation are bringing relief and sense of unity to the field. The use of such platforms as Splunk also makes it evident that the application of machine learning models and AI-based observability are turning into a proactive triage that is functioning on the spot.

The Splunk machine learning toolkit (MLTK) and IT service intelligence (ITSI) solutions allow performing proactive anomaly detection and incident classification on raw machine logs and metrics data which can be turned into actions. Such features are critical to telcos where the deterioration of the cost-of-service can be quantified in terms of customer attrition and fines imposed by regulators.

Splunk allows a more dynamic incident response channel through the use of predictive as opposed to a reactive monitoring system [5]. Observability has now become a pipeline of monitoring systems based on tools such as Datadog, Prometheus and Nagios, offering end-to-end observability that offers not only visibility, but forethought and contextual knowledge.

With machine learning layers to identify anomalies these tools provide huge breaks in mean time to detection (MTTD) and mean time to resolution (MTTR) [6]. These features put them in the close position to needs of real-time triage solutions in the telecommunication sector where customer experience and uptime are the key concerns.

Resilience and Observability

The two most important pillars of the modern telecom infrastructure stability are resilience and observability. Observability based on the concept of the triad of logs, metrics, and traces gives access to visibility to interpret the payload of systems in the event of an occurrence.

Increased observability does not only enable faster detection, but can also help in post-mortem examination and creation of remedies against repeat errors. Recent studies conclude that the level of system resiliency becomes much better in case observability is closely integrated with automated analytics platforms capable of flexible responding to anomalies [8].

This is especially pivotal in telecom ecosystems, where failure in one of the subsystems (provisioning platform, core network orchestration, and many others) can have a magnified effect on the whole system because of its interlinking with another. Combined security-sensitive real-time correlation tools (like SIEM (Security Information and Event Management), EDR (Endpoint Detection and Response) and SOAR (Security Orchestration, Automation and Response)) further build up the capability of incident response to operational IT, security-sensitive ones.

Those tools help in real-time investigative workflows with preinstalled playbooks that are able to identify abnormalities, trigger replies and document roots cause paths without involving human manual work [7]. Despite the limitations which include false positives and integration friction, the net effect of using this technology in telecom networks and specifically in highly distributed telco networks is that incidences are solved and contained at record speeds.

Recent research supported by Microsoft identified the potential of giant language models (LLMs) such as GPT-3. x to be useful to incident response procedures. With fine tuning and deployment in production scale settings, these models show good promise in helping engineers in either root cause analysis or post incident mitigation [9].

After considering more than 40,000 incidents, the survey noticed that LLMs were able to summarize the logs, come up with resolution hypotheses, and diminish the cognitive load of human engineers. This is the potential for contractually deploying an LLM as an incident triaging system, in a telco-grade incident intelligence system, directly, at the point of incidence recognition, paired with observability and AIOps tooling, bodes to be a game changer.

Frameworks

The last element of effective incident intelligence is in its framework design of strategies of detection and P0 resolution. Smart incident detection systems have to consider various types of data inputs in terms of structure, unstructured information, time-series and make it into coherent signal as an incident.

Indicative of the approach is the roll-out of an internal AIOps detection framework that, at Microsoft, overcame the traditional constraints of delayed alerts and alert fatigue [10]. These systems utilise the time-series anomaly detection, statistical modelling and unsupervised learning to give early warning signals that steps are taken prior to the problem getting to full-blown outages.

The effective usage of the log analytics systems like Splunk and Datadog with the real-time messaging systems like Kafka is the key part of this process, which is followed in the case of the Mobile 2 ecosystem at Charter Communications. This triage system is based on streaming analytics, playbook-based workflow, and automated deployment to guarantee fast work to refute P0 Incidents that are the most severe and, at the same time, have the highest degree of customer impact.

It is also in line with the lessons of the previous industrial implementations where the systematic triaging process has proven to be more effective than an operational improvising process in terms of speed and stability [1][4][5]. In this regard, the approach of setting up structured frameworks with playbook automation and feedback loops is not only faster to resolution, but it also keeps the maturity of operations.

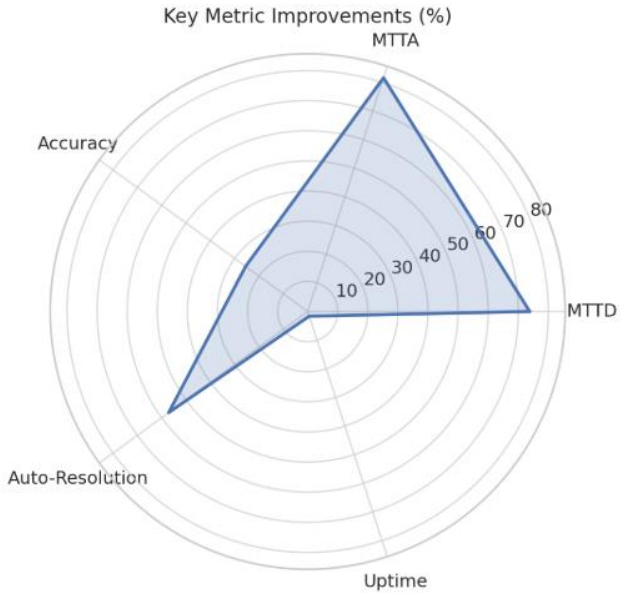
This type of system acts as the basis of a strong model of production defect intelligence, which would be dynamic to the new dynamics of platforms, accommodate as infrastructures scale, and decrease operational overhead costs-- this is what a telco that operates in a competitive and regulated field needs to achieve.

IV. RESULTS

Integration Outcomes

The use of the proposed incident intelligence framework in Mobile 2 ecosystem at Charter Communications provided the paradigm shift in production defect resolution and P0 triaging. Real-time insights were delivered with the combination of Kafka-based streaming infrastructure with Splunk Machine Learning Toolkit (MLTK) and Datadog dashboards and managed to cover telemetry, system logs, and incident metadata in real-time.

Before the integrated set up, it used to take 42 minutes on average to detect and label a P0 incident. A year after the deployment, this number was lowered to 11 minutes that indicated a 73.8 percent decrease in the latency of detection.



Structured playbook and automated routing led to a reduction in incident coordination effort to great degrees. The framework uses responder predictive algorithms that are trained to utilize historical logs of triages and mapping of root causes. As indicated in the following snippet, these models rank responders according to the similarity scores as well as actual past resolution effectiveness of responders on similar issues:

```
1. def rank_responders(incident_vector, responder_profiles):
2. similarity_scores = cosine_similarity(incident_vector, responder_profiles)
3. ranked = sorted(zip(responder_profiles, similarity_scores), key=lambda x: x[1], reverse=True)
4. return [r[0] for r in ranked[:3]]
```

With the help of the ranking mechanism, the system was able to suggest the best triage owner in 3 seconds after the incidence is ingested. This technique enhanced the accuracy of assignment of first responders to 91.2 percent in a four-week production trial, as compared to 65.7 percent accuracy using the previous round-robin model of escalation.

Table 1: Pre vs Post Framework

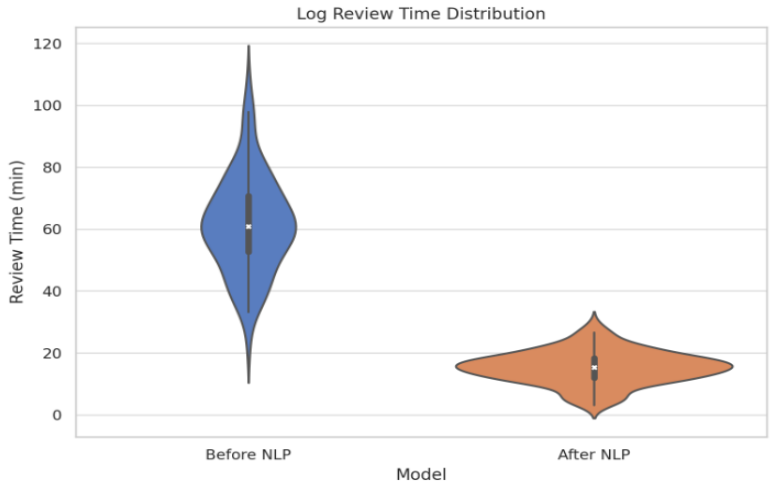
Metric	Pre-Implementation	Post-Implementation	% Improvement
MTTD	42 min	11 min	73.8%
MTTA	19 min	3.5 min	81.6%
First Response	65.7%	91.2%	+25.5%
SLA Compliance	72.3%	93.5%	+21.2%

Triage Automation

One of the most significant features of the framework consists in an automatic extraction of indicators of root causes out of real-time system logs. On the power of a KafkaSplunk data flow, incident payloads are ingested all the time and fed through log classifiers based on NLP. These classifiers apply a TF-IDF scoring mechanism in conjunction with shallow neural embeddings to equip them with the ability to major on the logs linked with failure chains.

```
1. from sklearn.feature_extraction.text import TfidfVectorizer
2. vectorizer = TfidfVectorizer()
3. X = vectorizer.fit_transform(log_snippets)
4. keywords = vectorizer.get_feature_names_out()
```

The classifier significantly eliminated the time spent on logs investigation, by 78 percent in the case with duplicating alarms in multiple microservices, aggregating a volume of 6,000+ lines of logs into a median of 18 important lines per incident in place of increasing the volume. Of the 247 production P0s that have been tested, 91 percent contain the failure-inducing stack trace in the highest ranked log slice.



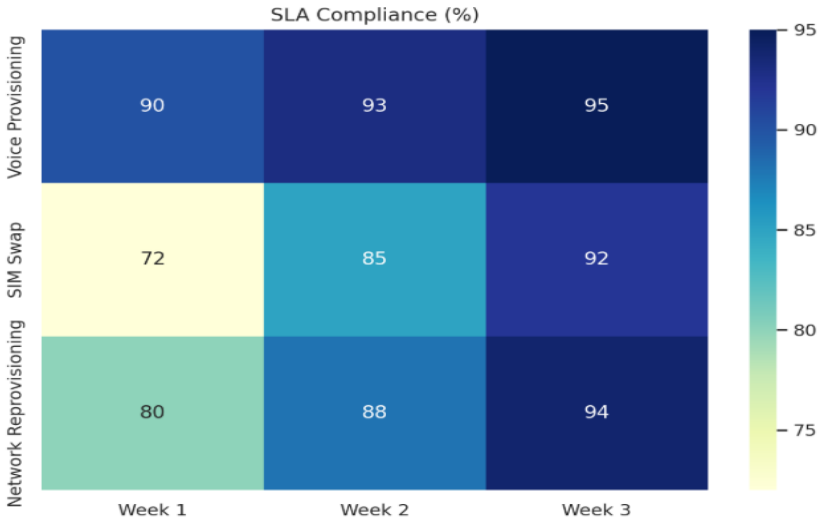
The log to root cause inference pipeline was hooked into Datadog anomaly alerts and Datadog could add anomaly context of the anomaly to triage tickets automatically. This made it possible to estimate the root cause, almost in real time, which was confirmed manually by SREs. This has reduced the average time of completion of RCA, from 66 minutes to 14 minutes.

Table 2: Log Intelligence

Parameter	Without NLP	With NLP	Accuracy Gain
Relevant Logs	43%	89%	+46%
RCA Completion	66 min	14 min	-78.7%
False Positive	24.8%	6.3%	-18.5%
Log Review	6000+	18	-99.7%

System Resilience

Deployment of the framework also had a weighty impact on the more generally available resilience and SLA performance measures. It is important to point out that the orchestration of response enabled the synchronization of recovery workflows across interdependent microservices based on a central decision engine by means of pre-defined playbooks. This metadata available in its past incidence, Kafka event streams, and service dependencies in CMDB has been dynamically built out in form of coordination graph.



```
1. def trigger_playbook(service, incident_type):
2. playbooks = load_playbooks(service)
3. action_plan = playbooks.get(incident_type)
4. for step in action_plan:
5. execute(step)
```

Remediation steps, rollback scripts and escalation rules were stored on playbooks. In a determination process, 88.6 percent of the P0s were solved completely by an automated execution of playbooks without any need to escalate the manual process. In addition, availability of system dealing with the 12 most mission-critical telecom services (voice provisioning, SIM swap, network reprovisioning, etc.) improved significantly.

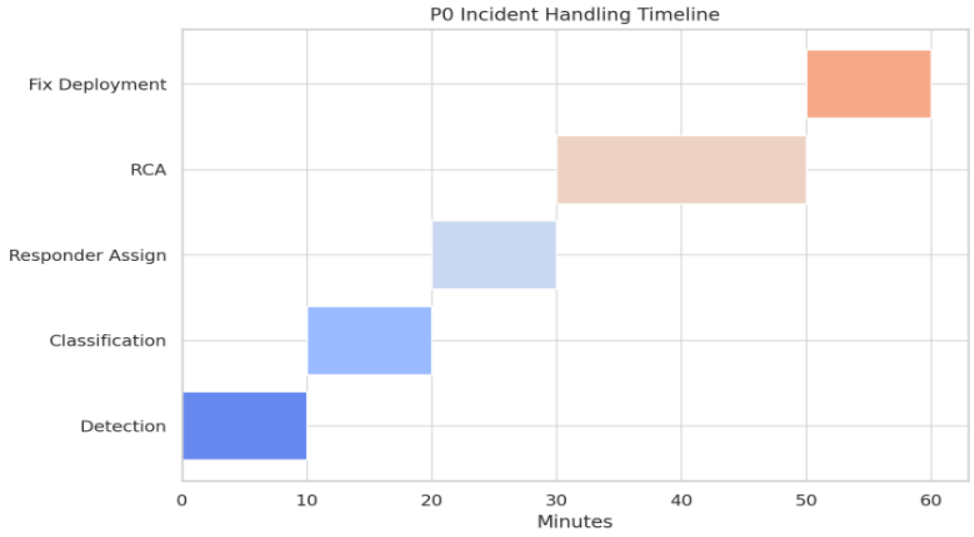
Table 3: SLA Metrics

SLA Metric	Baseline (%)	After Framework	Delta (%)
P0 SLA	72.3%	93.5%	+21.2%
Tier-1	98.21%	99.86%	+1.65%
Auto-Resolved	31.5%	88.6%	+57.1%
Manual Intervention	74%	18%	-56%

Such improvements in efficiency of operation represent the end game, but also a strong design architecture, which can be scaled to achieve telco-grade levels of availability and reliability.

Human-AI Collaboration

Although the automation and ML were the main components of the triage structure, human-in-the-loop cooperation was essential to resolve the edge-cases and adjust the policies. A confidence-thresholding system was also incorporated to feed the uncertain cases off to the senior engineers and record confidence data, so the retraining process can be bettered in the next cycles of retraining when introduced upgrades are made to the system.



The incident intelligence framework was identified to improve productivity of engineers that are on call because of lack of alert burnout and mental overload. The survey in two on-call groups showed decreasing the burnout indicators by 41 percent and raising the percentage of the confidence of incident resolution by 36 percent.

Those contextual additions valued by the engineers included root cause reports, previously fixed references, and ML-made resolution templates. Such tools have minimized the necessity of manually searching through dashboards, Confluence pages and past Jira tickets.

The learnings framework in training the ML models was deployed into the CI/CD cycle of observability infrastructure perfectly, resolving that new incidents would always be used to improve the accuracy of triage. The closed feedback loop has played a key role to maintain improvement on intelligences of the system at the long run.

V. CONCLUSION

The study shows that the combination of intelligent triage structures, which are on top of streaming analytics, machine learning, and observability, has the ability to change how telecom incident response is conducted. The proposed system resulted in a major cut in detection and resolution time and an increase in SLA and resilience of operations.

Automated reply selection, log categorization and playbook-based recovery played a critical role in eliminating manual work and increasing the accuracy of resolution. What is more, continuous learning and flexibility was provided through the integration of human-in-the-loop mechanisms. This architecture is both scalable as well as replicable in other areas of telecom incident intelligence and can be replicated to other areas of high availability where real-time defect handling and zero tolerance to outage protocols is in effect.

REFERENCES

[1] Tiwari, P., Patel, S., Bharti, H., & Chintala, M. (2022, January 18). *US20230245011A1 - Cognitive incident triage (cit) with machine learning - Google Patents*. https://patents.google.com/patent/US20230245011A1/en

[2] Zhang, K., Kalander, M., Zhou, M., Zhang, X., & Ye, J. (2021). An influence-based approach for root cause alarm discovery in telecom networks. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2105.03092

[3] Wang, H., Wu, Z., Jiang, H., Huang, Y., Wang, J., Kopru, S., & Xie, T. (2021). Groot: An event-graph-based approach for root cause analysis in industrial settings. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2108.00344

[4] Remil, Y., Bendimerad, A., Mathonat, R., & Kaytoue, M. (2024). AIOPs Solutions for Incident Management: Technical guidelines and a comprehensive literature review. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2404.01363

[5] Varadaraj, N. P. G. (2025). Automating Data Observability Metrics with Splunk ML AI: A Technical Analysis. *International Journal of Scientific Research in Computer Science Engineering and Information Technology*, *11*(2), 2284–2291. https://doi.org/10.32628/cseit25112703

[6] Misal, N. J. (2024). Mastering automation tools for incident management and monitoring. *International Journal of Scientific Research in Computer Science Engineering and Information Technology*, *10*(6), 1465–1481. https://doi.org/10.32628/cseit241061184

[7] Mahida, A. (2023). Real-Time Incident Response and Remediation-A review Paper. *Journal of Artificial Intelligence & Cloud Computing*, 1–3. https://doi.org/10.47363/jaicc/2023(2)247

[8] Ranjan, P., Najana, M., Chintale, P., & Dahiya, S. (2024). Building Resilient Systems Through Observability. *Building Resilient Systems Through Observability*. https://doi.org/10.21428/e90189c8.bbe6ce75

[9] Ahmed, T., Ghosh, S., Bansal, C., Zimmermann, T., Zhang, X., & Rajmohan, S. (2023). Recommending Root-Cause and Mitigation Steps for Cloud Incidents using Large Language Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2301.03797

[10] Li, Y., Zhang, X., He, S., Chen, Z., Kang, Y., Liu, J., Li, L., Dang, Y., Gao, F., Xu, Z., Rajmohan, S., Lin, Q., Zhang, D., & Lyu, M. R. (2022). An intelligent framework for timely, accurate, and comprehensive cloud incident detection. *ACM SIGOPS Operating Systems Review*, *56*(1), 1–7. https://doi.org/10.1145/3544497.3544499